Hands-on tutorial on data assimilation using Markov state model and Hidden Markov models

Yasuhiro Matsunaga

ymatsunaga@mail.saitama-u.ac.jp

What can we learn?

- We will learn basics of integrating experimental data to MD simulation data using Markov state model (MSM) and Hidden Markov state model (HMM)

- Starting from an introduction to our Julia package (MDToolbox.jl), we will learn basics of MSM and HMM using simple examples. Then, we will combine MSM and HMM for incorporating experimental data.

Schedule

- 13:00 Installation. Julia and MDToolbox.jl basics
  - Introduction to our in-house Julia package MDToolbox.jl
- 14:00 Markov state model (MSM) basics
  - What is MSM?, and how to construct it
- 14:40 Hidden Markov model (HMM) basics
  - What is HMM? and what can we do with it
- 15:20-16:00 Data assimilation with MSM & HMM

00 Installation. Julia and MDToolbox.jl basics

- Let's follow the descriptions of README at https://github.com/matsunagalab/hmm_tutorials

1. Install Jupyter Lab
2. Install Julia and packages
    1. Install Julia
    2. Install required packages including MDToolbox.jl (or update)
    3. Run Ijulia
3. Download tutorial materials

MDToolbox.jl

Julia package for the statistical analysis of MD trajectories

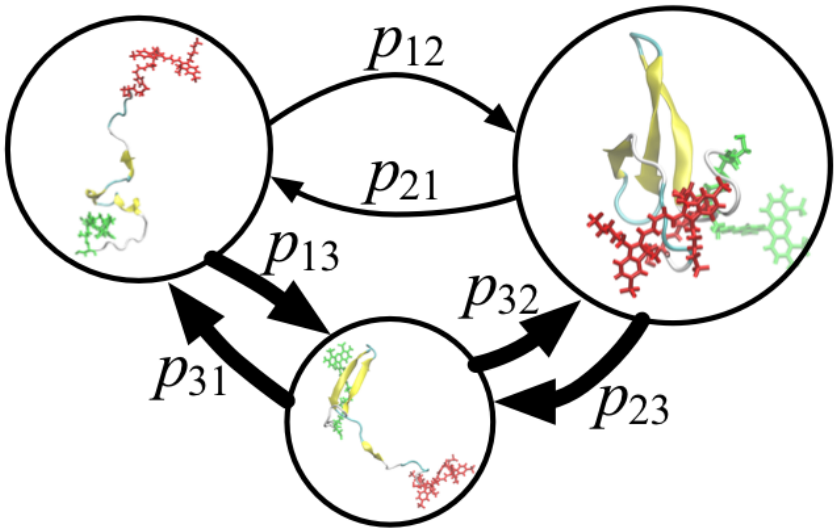
https://github.com/matsunagalab/MDToolbox.jl

- I/O for trajectory, coordinate, and topology files
- Atom selections
- Least-squares fitting of structures, and some geometry calculations
- Potential mean force (PMF) or free energy profile from scattered data
- Statistical estimates (WHAM and MBAR methods) from biased data
- Dimensional reductions (Principal Component Analysis, and others)
- Markov state models and Hidden Markov models

01 Markov state model basics

What is MSM? MSM is a statistical model to describe conformational dynamics of molecules

Structures are clustered into **discrete states**



Dynamics are represented by **transition probabilities** between states

$$T_{\text{simulation}}(\tau) = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

Time is coarse-grained in τ

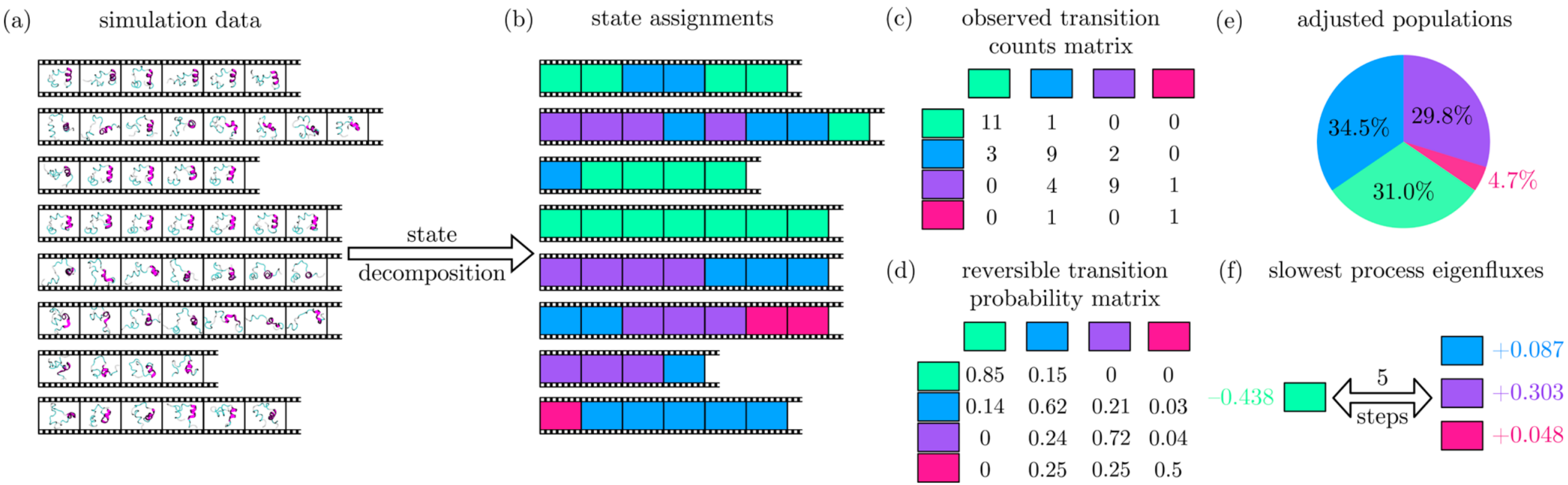
$$P(n\tau) = \left[T_{\text{simulation}}(\tau)\right]^n P(0)$$

🤗 Dynamics is represented by parameters $T(\tau)$

🤗 Long time dynamics from multiple short simulations

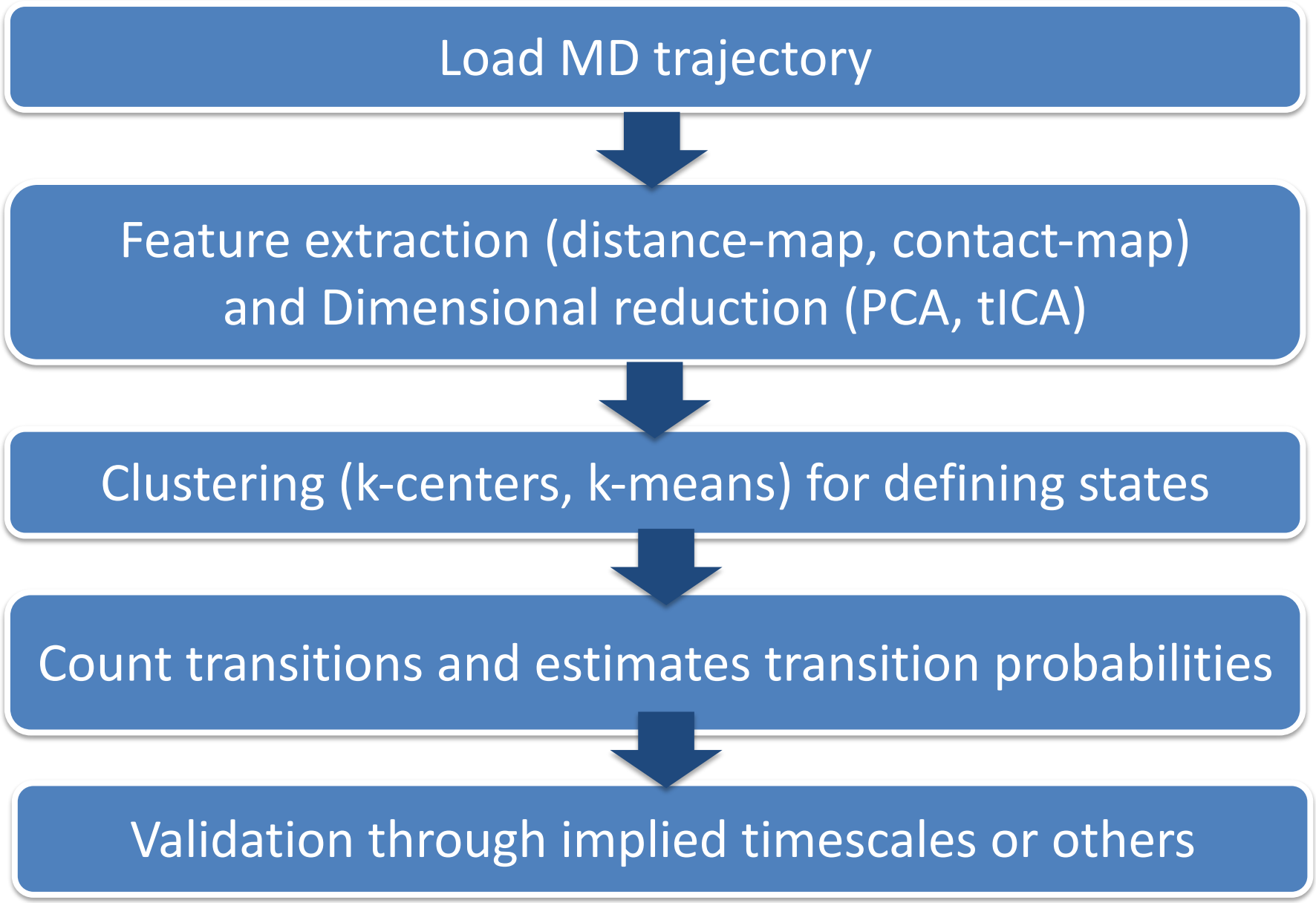
😞 Force-field dependence (Incorrect dynamics)

Construction of MSMs



B. E. Husic, V. S. Pande, Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **140**, 2386–2396 (2018).

Typical computations for constructing MSMs



MDToolbox functions

`mdload()`

`compute_distancemap()`

`pca()`

`tica()`

`clusterbykcenter()`

`msmcountmatrix()`

`msmtransitionmatrix()`

`msmimpliedtime()`

Estimation of transition probabilities from counting matrix: Maximum likelihood

$$L_{\mathrm{MD}}\left(\mathbf{T}(\tau)\right) = p\left(\mathbf{C}|\mathbf{T}(\tau)\right) = \prod_{i=1}^{M}\prod_{j=1}^{M} T_{ij}^{C_{ij}}(\tau).$$

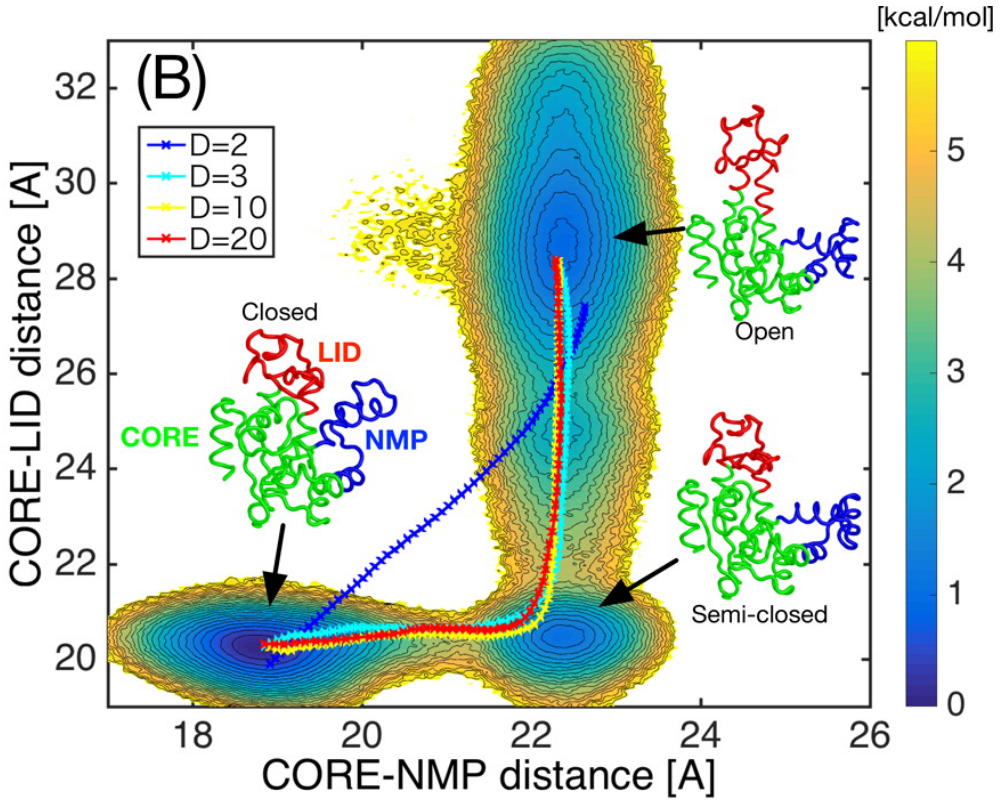
Maximize this function over $T(\tau)$

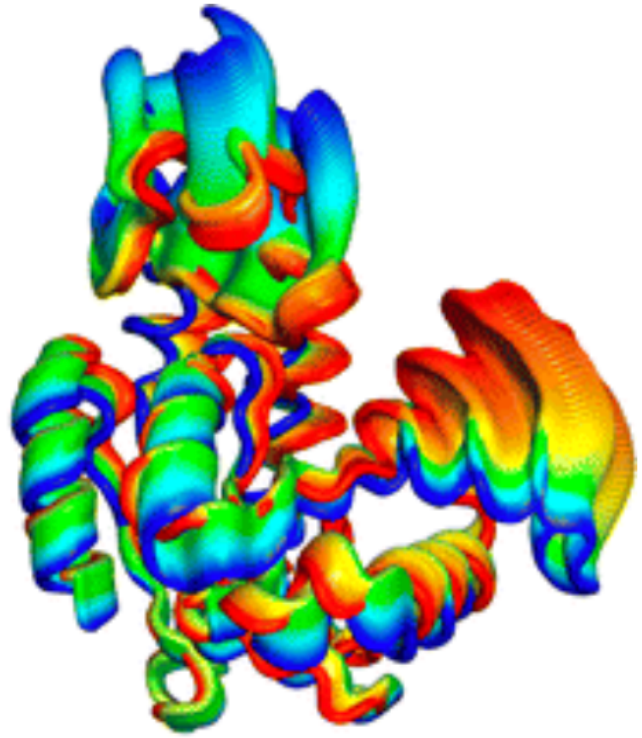
K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, V. S. Pande, MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **7**, 3412–3419 (2011).

Y. Matsunaga, Y. Sugita, Use of single-molecule time-series data for refining conformational dynamics in molecular simulations. *Curr. Opin. Struct. Biol.* **61**, 153–159 (2020).

MD data in the tutorial

Coarse-grained (DoME-model) model of Adenylate kinase, simulated by GENESIS





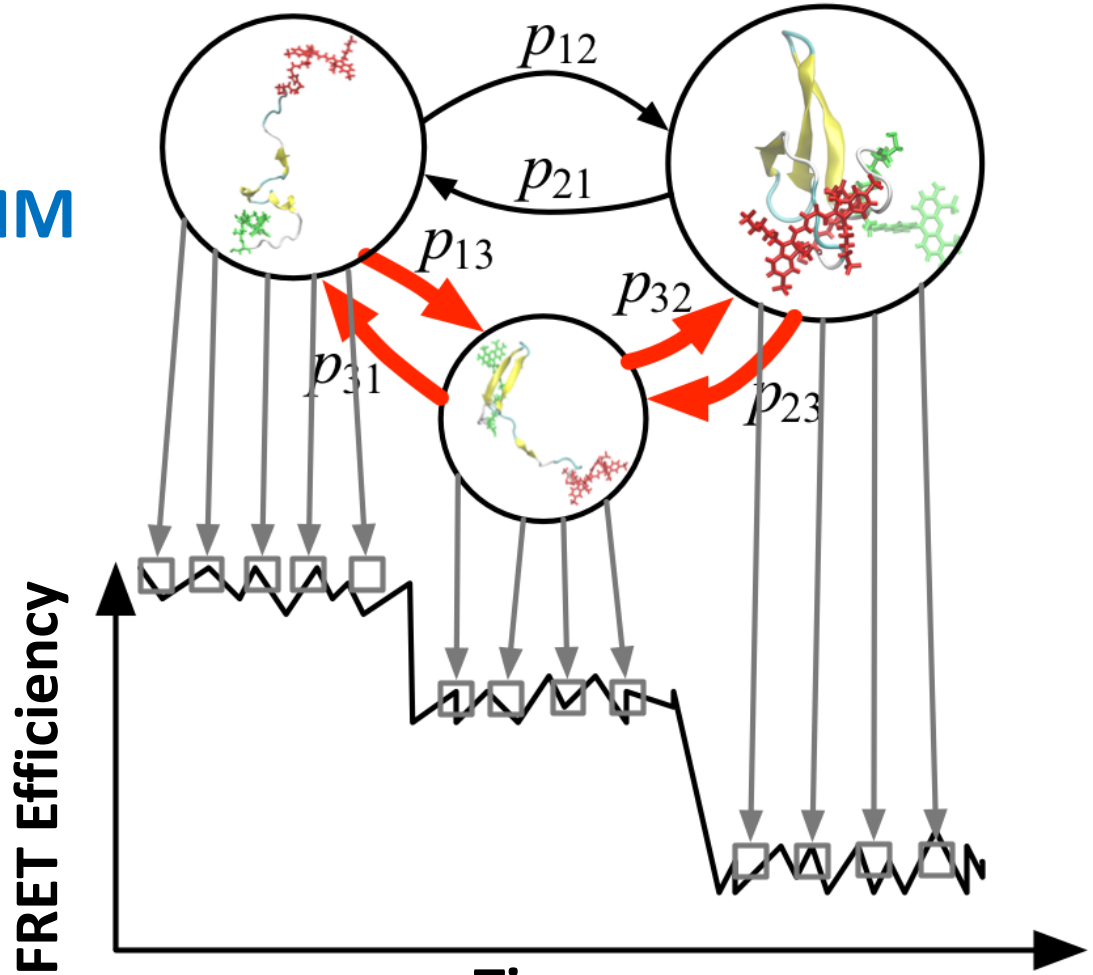
C. Kobayashi, Y. Matsunaga, R. Koike, M. Ota, Y. Sugita, Domain Motion Enhanced (DoME) Model for Efficient Conformational Sampling of Multidomain Proteins. *J. Phys. Chem. B*. **119**, 14584–14593 (2015).

Y. Matsunaga, Y. Komuro, C. Kobayashi, J. Jung, T. Mori, Y. Sugita, Dimensionality of Collective Variables for Describing Conformational Changes of a Multi-Domain Protein. *J. Phys. Chem. Lett.* **7**, 1446–1451 (2016).

02 Hidden Markov model basics

What is HMM? MSM is a latent MSM behind experimental observations

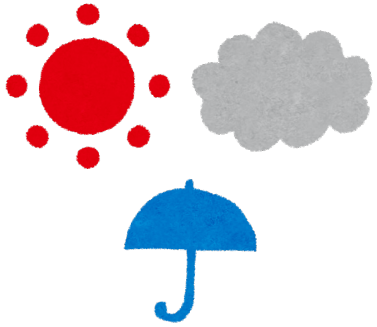
HMM



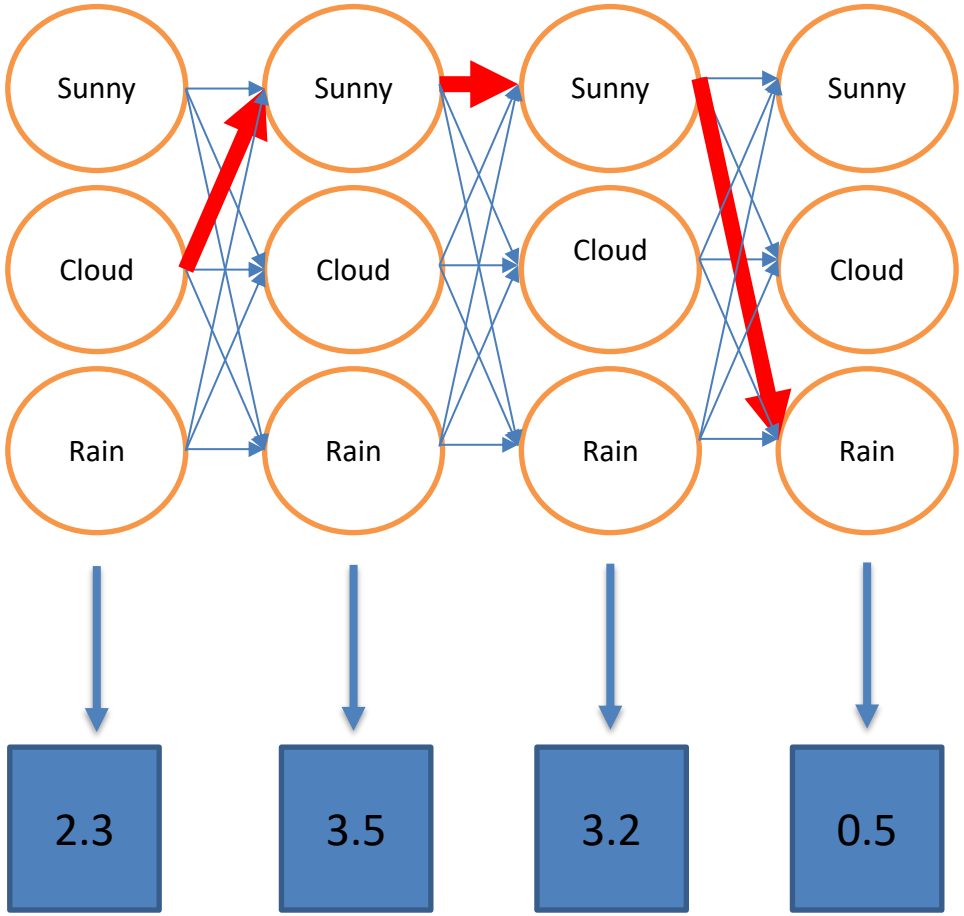
Experimental observations

Toy model used in the tutorial

Daily weathers



Daily Ice-cream sales



Hidden Markov Model

Can we estimate daily weathers Or transition probabilities from daily ice-creams sales Time-series data?

Experimental observations

Toy model used in the tutorial (cont'd)

Transition probabilities of daisy weathers

Weather	value
Rain	x = 1
Cloudy	x = 2
Sunny	x = 3

	Rain (tomorrow)	Cloudy(tomorrow)	Sunny (tomorrow)
Rain (today)	0.200004	0.257922	0.542074
Cloudy (today)	0.242095	0.300014	0.457891
Sunny (today)	0.157926	0.142121	0.699953

Probabilities to observe specific daily sales values given weathers

$$Y = X + \epsilon$$

sales weathter Gaussian noise

$$\epsilon = Y - X$$

$$P(y|X = x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-x)^2}{2} \right)$$

sales weathter

Estimation of latent states and transition probabilities from counting matrix: Maximum likelihood

$$L_{\mathrm{HMM}}\left(\mathbf{T}(\tau)\right) = p\left(y_{1:N}|\mathbf{T}(\tau)\right)$$

$$= \sum_{s_1=1}^{M} \cdots \sum_{s_N=1}^{M} p(s_1)p(y_1|\mathrm{s}_1) \prod_{t=2}^{N} T_{s_{t-1}s_t}(\tau)p(y_t|s_t).$$

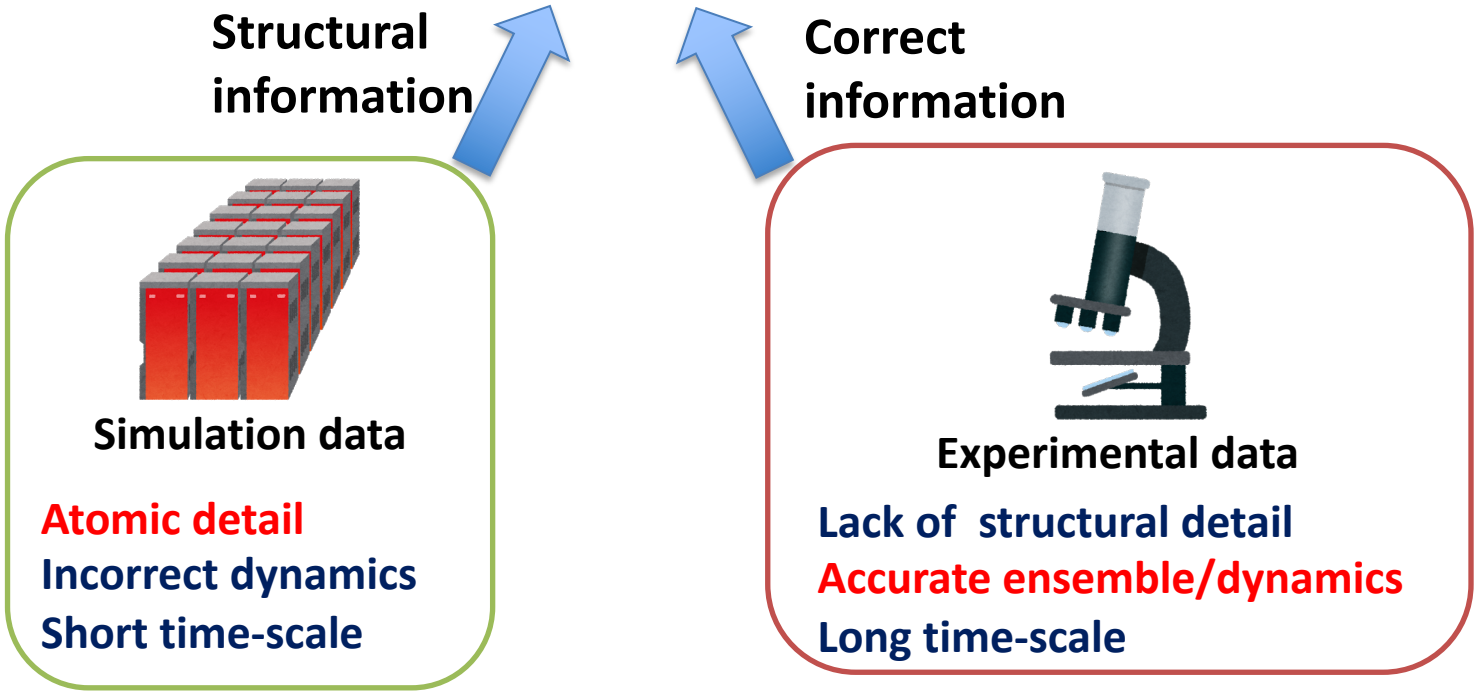
Maximize this function over states or $T(\tau)$

Y. Matsunaga, Y. Sugita, Use of single-molecule time-series data for refining conformational dynamics in molecular simulations. *Curr. Opin. Struct. Biol.* **61**, 153–159 (2020).

03 Data assimilation with MSM and HMM

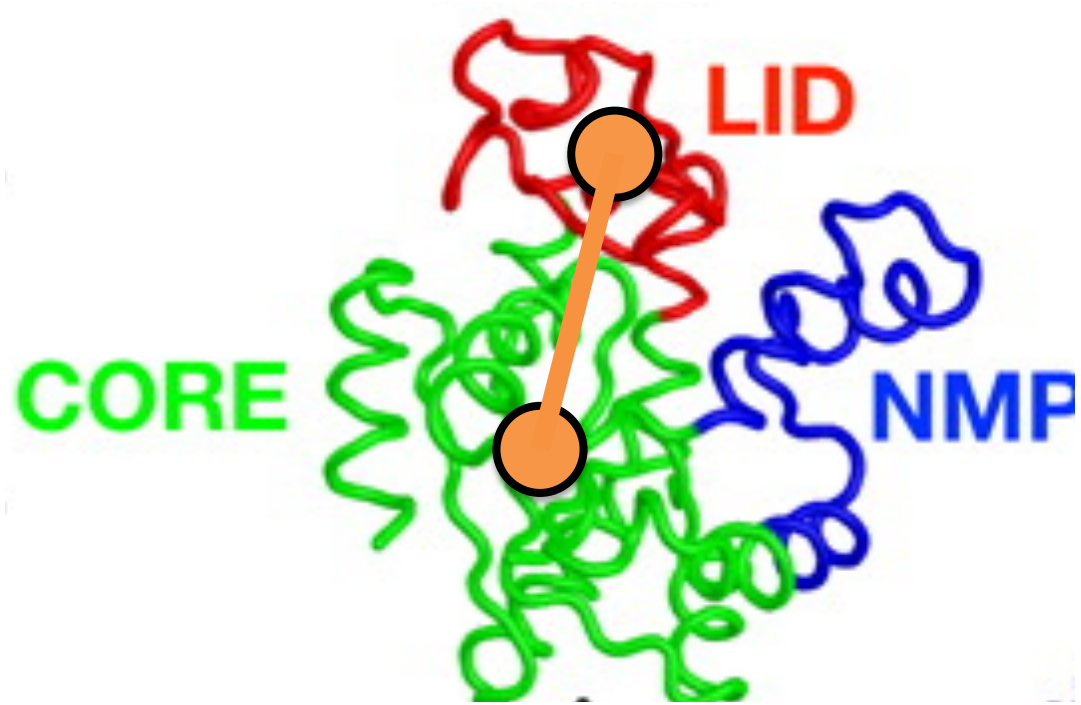
What is the purpose of integrating MD data and HMM data?

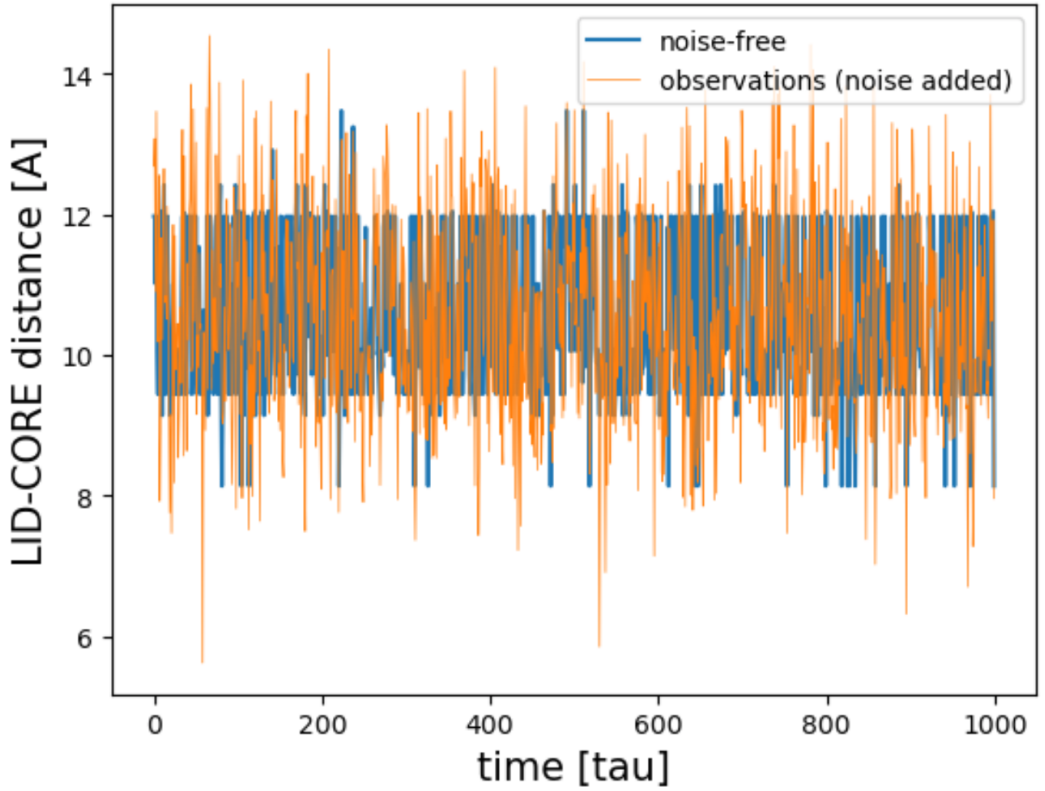
- **Correct MD data ensemble or dynamics matching with experimental data**
- **Interpret experimental data in terms of structural details**



”Experimental data” in the tutorial

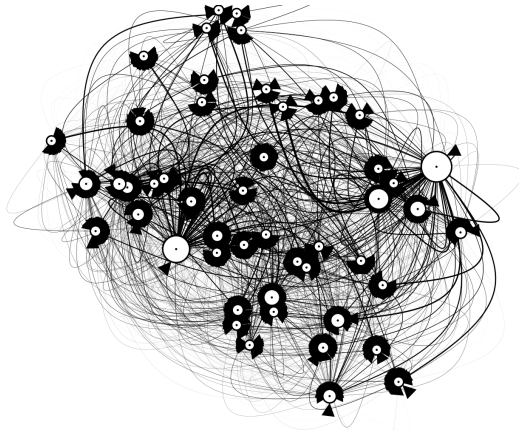
Using the original MSM, we generate LID-CORE distance time-series as ”experimental data” mimicking single-molecule FRET





”Experimental data” in the tutorial (cont’d)

Transition probabilities of MSM’s states



Probabilities to observe specific daily sales values given weathers

$$Y = X + \epsilon$$

observation: Y; LID-CORE distance: X; Gaussian noise: ϵ

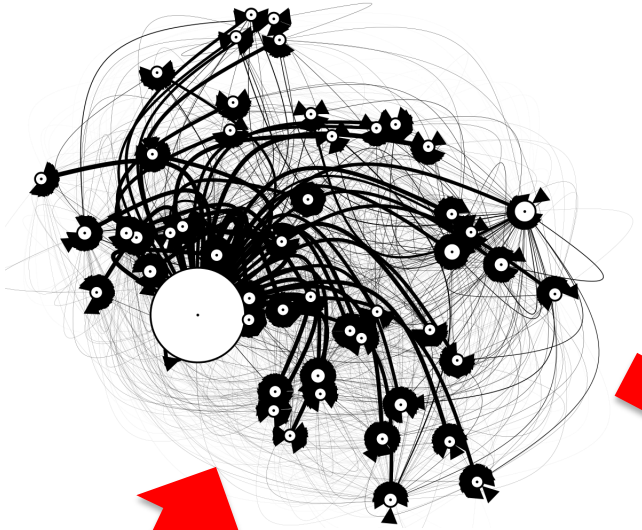
$$\epsilon = Y - X$$

$$P(y|X = x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - x)^2}{2} \right)$$

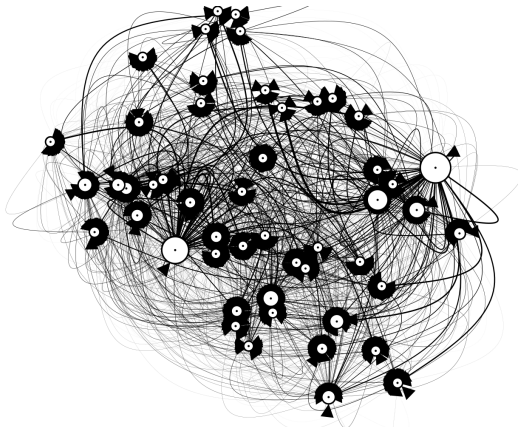
observation: y; LID-CORE distance: x

Problem setting in the tutorial

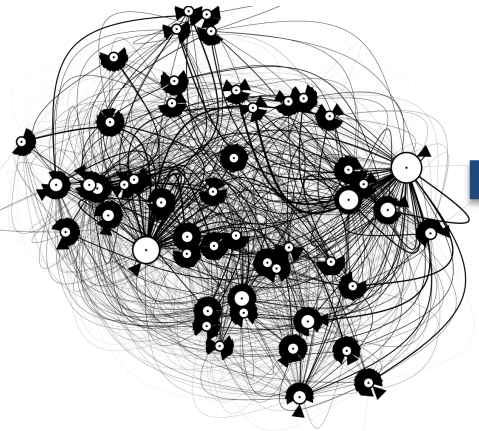
"Distorted" MSM

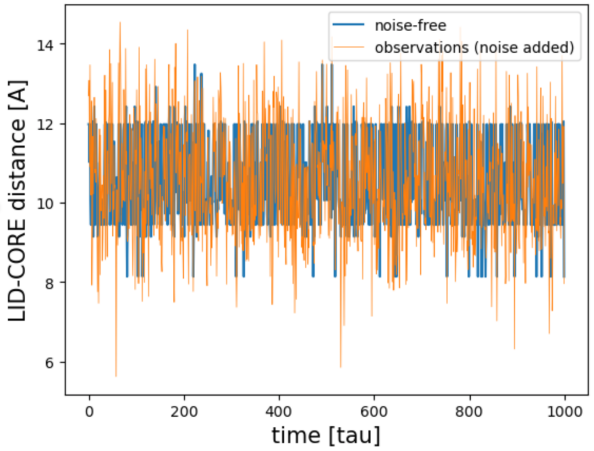


Can we correct the "distorted" MSM and restore the original MSM?



"experimental data" generated by the original MSM





Data assimilation or integration