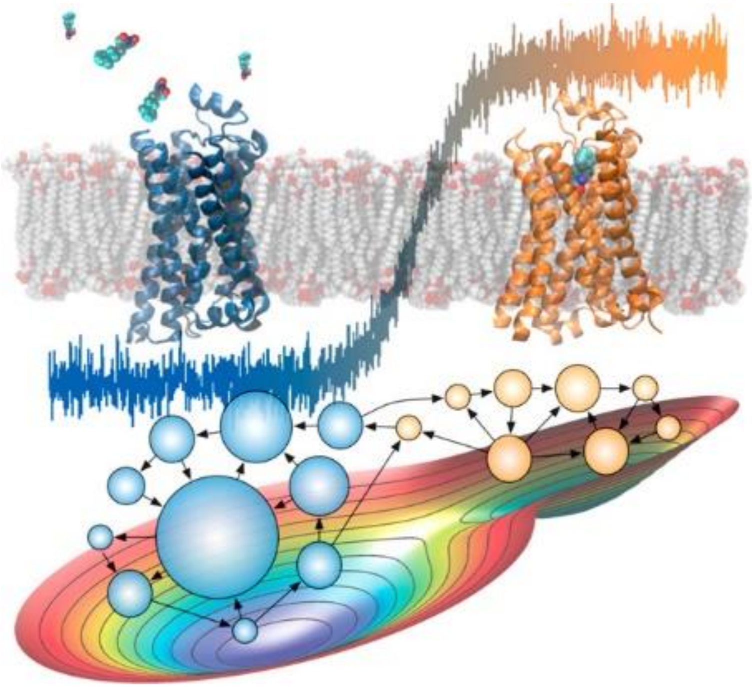
Thermodynamics + Kinetics
- Markov state modelings



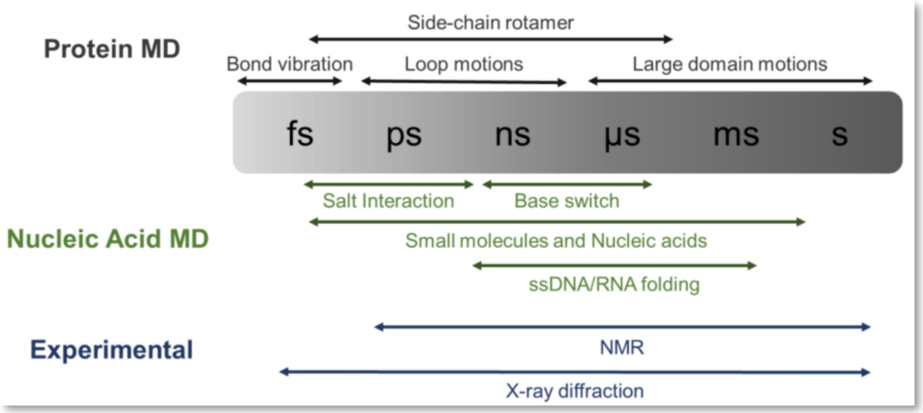
2024 Winter Son Lab Seminar
February 23th, 2024

Junho Lim

[1] *Acc. Chem. Res.* 2015, 48, 2, 414–422

Appetizer : We always think the timescale of dynamics

■ Biological process has the long range of timescales.



[1] *Biomolecules* **2018**, 8, 83.

■ Many sampling methods can explore the phase space efficiently. But, to investigate the “kinetics”, we need “Time axis!”

■ Markov State Modelings(MSMs) can bridge this timescale gap by modeling the long timescale dynamics based on many short MD simulations.

Then, Let’s ask. **1) What’s the meaning of ‘Markov’?**
2) How do we set MSMs?
3) What are the applications & challenges for MSMs?

Content

Introduction) Markov chain : Memoryless

Building MSM : How do we partition the space and time?

Analysis MSM : What quantities can be calculated?

Further MSM : Wake up! It’s time for math.

Content

Introduction) Markov chain : Memoryless

Building MSM : How do we partition the space and time?

Analysis MSM : What quantities can be calculated?

Further MSM : Wake up! It’s time for math.

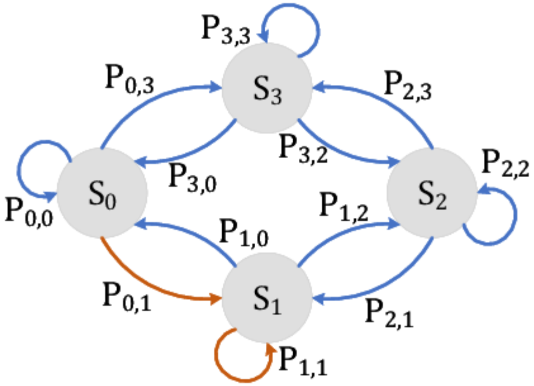
Introduction to Markov chains : Memoryless

■ The most important keyword you should remember in Markov chains is **"Memoryless"**

■ Def] Markov Process : A stochastic process where the future state only depends on the present state and all the past states are eliminated.

Let's consider the (discrete) Markov chains.

x_k $(k = 0,1,2, \dots)$: a random variable, mapping into a finite state space $S = \{S_1, \dots, S_n\}$.



■ Markov process satisfies the memoryless property for all $k \geq 1$ and states $S_0, \dots, S_k$:

$$P(x_k = S_k | x_{k-1} = S_{k-1}, \dots, x_0 = S_0) = P(x_k = S_k | x_{k-1} = S_{k-1})$$

In short, we will write

$$P(x_k | x_{k-1}, \dots, x_0) = P(x_k | x_{k-1})$$

■ Def] Transition matrix : $T \in R^{n \times n}$: $T_{ij} = P(x_k = j \mid x_{k-1} = i)$

Properties of the transition matrix

1. $T_{ij} \geq 0 \ \forall i, j$
2. $\Sigma T_{ij} = 1 \ \forall i$

[The lecture notes will be given.]

Introduction to Markov chains : Memoryless

■ The most important keyword you should remember in Markov chains is **"Memoryless"**

When we think about the probability to find the chain at **state $\boldsymbol{i}$ at time k,**

$$p_{k,i} = p_{k-1,1}T_{1i} + \ \dots + p_{k-1,n}T_{ni} = \Sigma p_{k-1,j}T_{ji}$$

Define the probability vector $\boldsymbol{p}_k = \left(p_{k,1}, \dots, p_{k,n}\right)^T$, this is compactly written as :

$$\boldsymbol{p}_k^T = \boldsymbol{p}_{k-1}^T\boldsymbol{T}$$

Applying this equation k times : Chapman-Kolmogorov equation :

$$\boldsymbol{p}_k^T = \boldsymbol{p}_0^T\boldsymbol{T}^k$$

■ Def] A probability distribution $\pi \in R^n$ is a stationary distribution of $\boldsymbol{T}$ when :

$$\boldsymbol{\pi}^T\boldsymbol{T} = \boldsymbol{\pi}^T$$

(Note : $\pi\ exists\ and\ unique\ when\ the\ T\ matrix\ is\ irreducible\ and\ reversible. \rightarrow\ H.W.$)

Note : After we set up the transition matrix, we could calculate the stationary distribution of T!

[The lecture notes will be given.]

Content

Introduction) Markov chain : Memoryless

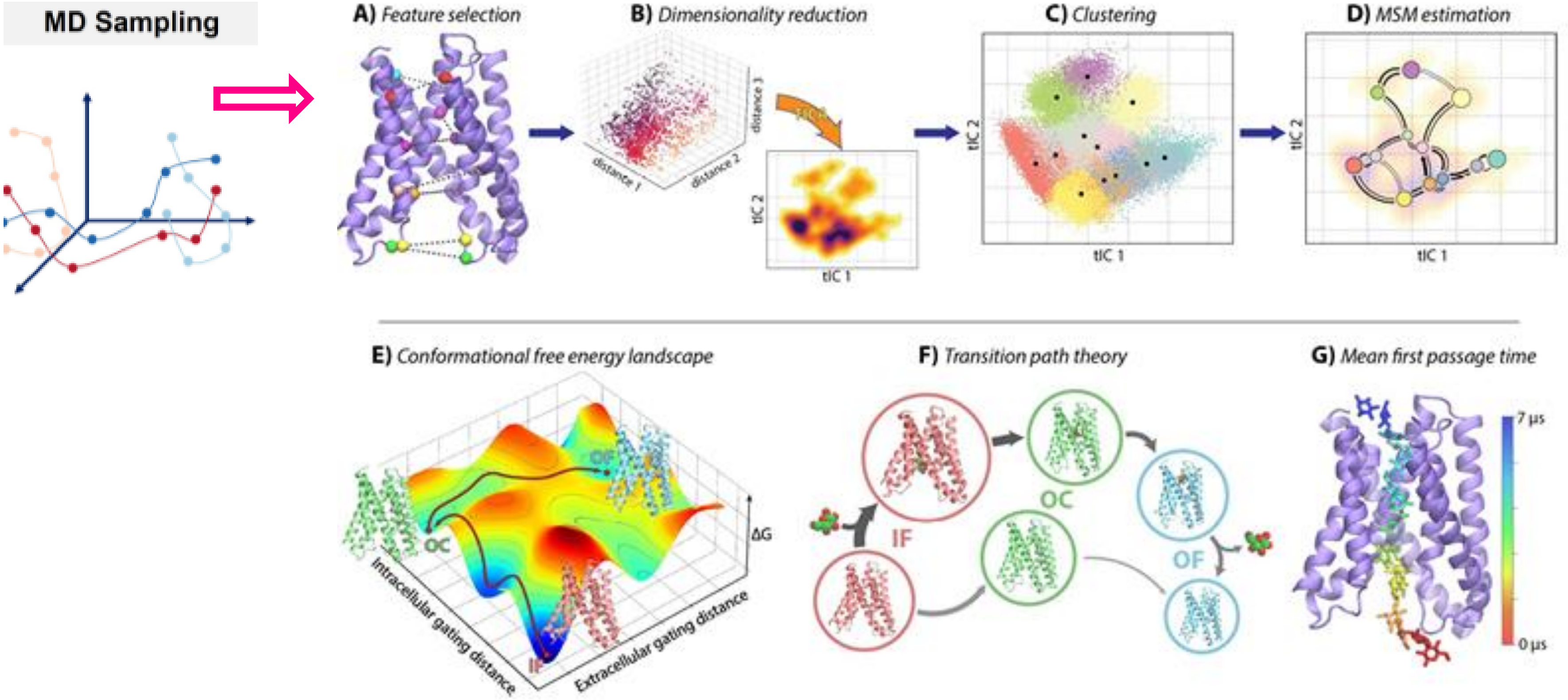
Building MSM : How do we partition the space and time?

Analysis MSM : What quantities can be calculated?

Further MSM : Wake up! It’s time for math.

How do we partition the space and time?

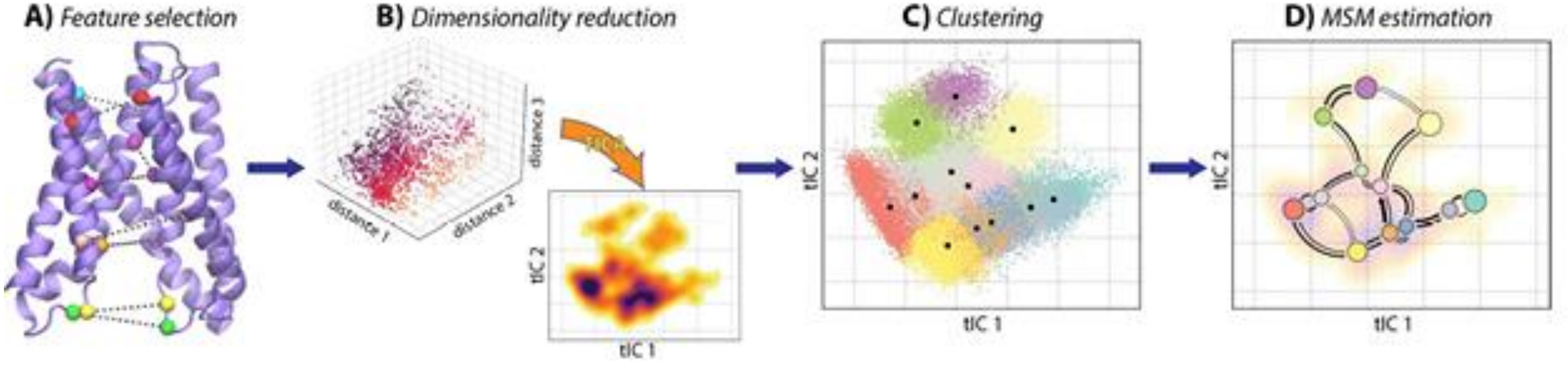
■ Let’s see the entire pipeline to set and run MSMs


MD Sampling
A) Feature selection
B) Dimensionality reduction
distance 1
distance 2
distance 3
tICA
tIC 2
tIC 1
C) Clustering
tIC 2
tIC 1
D) MSM estimation
tIC 2
tIC 1
E) Conformational free energy landscape
OF
OC
IF
ΔG
Intracellular gating distance
Extracellular gating distance
F) Transition path theory
IF
OC
OF
G) Mean first passage time
7 μs
0 μs

[1] *J. Struct. Biol.* **2021**, *213*, 107800

How do we partition the space and time?

■ MSMs - step 1 : Construction



- Run MD simulation
- Calculate internal coordinates (e.g. inter-residue distances)

- Reduce dimensionality to identify several CVs
- CVs describe the slowest dynamics of the system
- E.g.) PCA

- Partitioning the reduced-dimensional conformational space
- Centroid-based algorithms
- E.g.) K-means/Centers/Medoids

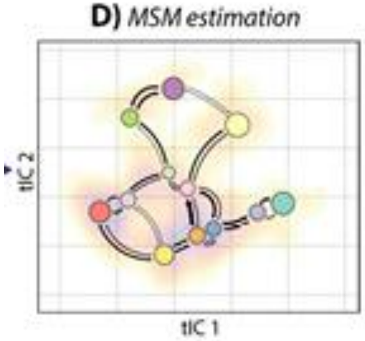
- Constructing Markov state Modelings
- Detailed balance & Maximum likelihood estimator (MLE)
- Described next.

[1] *J. Struct. Biol.* **2021**, *213*, 107800

[2] Springer Science & Business Media, **2013**, Vol. 797.

How do we partition the space and time?

■ MSMs - step 1 : Construction



- Constructing Markov state Models
- Detailed balance & Maximum likelihood estimator (MLE)
- Described next.

■ estimation of transition matrix :

$$T_{ij}(\tau) = p[x(t+\tau) \in j | x(t) \in i] = \frac{C_{ij}(\tau)}{\Sigma_j C_{ij}(\tau)}$$

(C : transition count matrix (TCM)
$C_{ij}(\tau)$: corresponds to the number of transitions that begin from state I and end at state j after the lag time τ)

■ Detailed balance :

$$\boldsymbol{C}^{sym}(\tau) = \frac{\boldsymbol{C}(\tau) + \boldsymbol{C}(\tau)^T}{2}$$

If there are large differences between $\boldsymbol{C}_{ij}(\tau)$ and $\boldsymbol{C}_{ji}(\tau)$: Use MLE

$$p\left(T\middle|C^{obs}\right) \propto \prod_{i,j=1}^{n} T_{ij}^{C_{ij}^{prior}+C_{ij}^{obs}} = \prod_{i,j=1}^{n} T_{ij}^{c_{ij}}$$

Results :

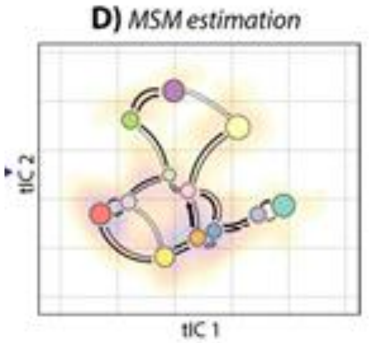
$$\pi_i = \sum_j \frac{C_{ij} + C_{ji}}{\frac{N_i}{\pi_i} + \frac{N_j}{\pi_j}}, \qquad T_{ij} = \frac{\left(c_{ij} + c_{ji}\right)\pi_j}{N_j\pi_i + N_i\pi_j}$$

[1] *J. Chem. Phys.*, **2011**, 134, 174105.

[2] Springer Science & Business Media, **2013**, Vol. 797.

How do we partition the space and time?

■ MSMs - step 1 : Construction



- Constructing Markov state Models
- Detailed balance & Maximum likelihood estimator (MLE)
- Described next.

■ Example :

For trajectory, the states : [1,1,2,2,2,1,2,1,2,1,2]

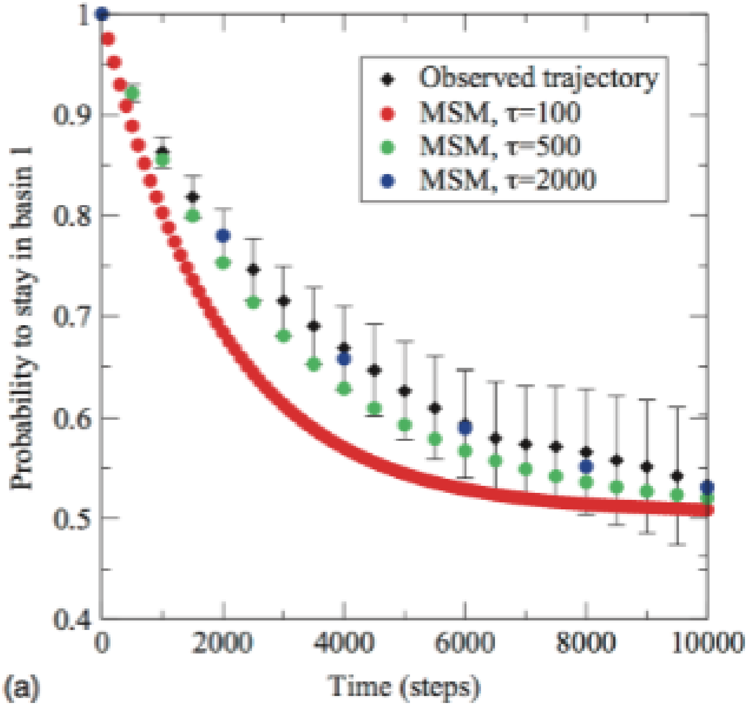
1) Transition count matrix (TCM) : $N_{11} = 1, N_{12} = 4, N_{21} = 3, N_{22} = 2 \rightarrow \begin{pmatrix} 1 & 4 \\ 3 & 2 \end{pmatrix}$
2) Detailed balance : $N^{symm} = \frac{N+N^T}{2} \rightarrow \begin{pmatrix} 1 & 3.5 \\ 3.5 & 2 \end{pmatrix}$
3) Generate TPM : $P_{ij} = \frac{N_{ij}^{symm}}{\Sigma(N_{ij}^{symm})} \rightarrow \begin{pmatrix} 0.222 & 0.778 \\ 0.636 & 0.364 \end{pmatrix} = \boldsymbol{T}$

■ After We set the TPM, we can do several analysis as shown below. Before that, Let’s validation our MSM model, basically, using Chapman-Kolmogorov equation!

Review : $\boldsymbol{p}_k^T = \boldsymbol{p}_0^T \boldsymbol{T}^k$

How do we partition the space and time?

■ MSMs - step 2 : Validation



■ The major validation of MSMs is the lag time.

■ Chapman-Kolmogorov Test : Using $\boldsymbol{p}_k^T = \boldsymbol{p}_0^T \boldsymbol{T}^k$

■ Check if our model shows Markovian property by checking

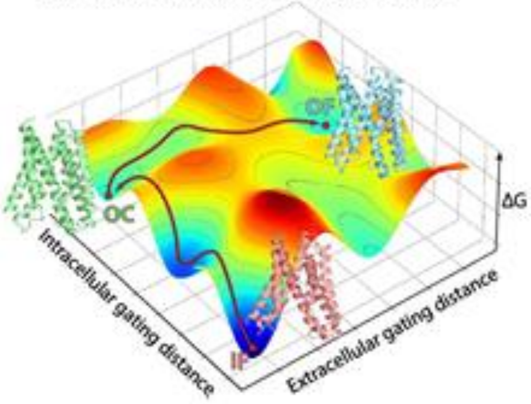
$$P_{MD}(n\tau) = [P_{MSM}(\tau)]^n$$

[1] *J. Chem. Phys.*, **2011**, 134, 174105.

How do we partition the space and time?

■ MSMs - step 3 : Basic analysis

E) *Conformational free energy landscape*



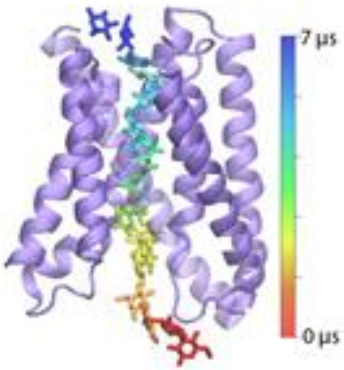
■ After we set & validate our transition matrix, we can calculate the stationary state!

■ With this stationary state, we can calculate the conformational free energy landscape

■ Thermodynamic quantity : The stationary state

-> Calculate π^T vector, which satisfies $\boldsymbol{\pi}^T\boldsymbol{T} = \boldsymbol{\pi}^T$ (eigenvalue problem)

G) *Mean first passage time*



■ We have the information of lag-time and the probabilities between each two states.

■ kinetic quantity : MFPT (Mean First Passage Time)

$$F_{if} = \tau + \sum_{j\backslash\neq f} P_{ij}F_{if}$$

Content

Introduction) Markov chain : Memoryless

Building MSM : How do we partition the space and time?

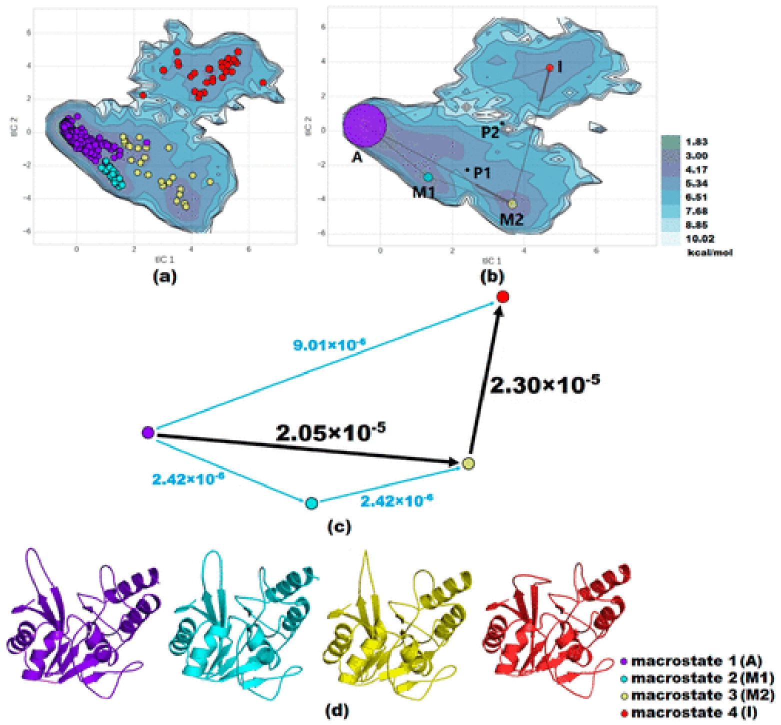
Analysis MSM : What quantities can be calculated?

Further MSM : Wake up! It's time for math.

Analysis MSM : Let’s see the fancy applications

■ Let’s see some references and check “What they calculated.”

Paper 1 : A Network of Conformational Transitions in the Apo Form of NDM-1 Enzyme Revealed by MD Simulation and a Markov State Model



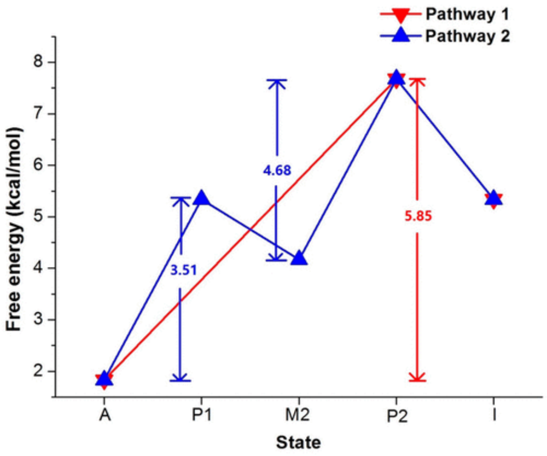
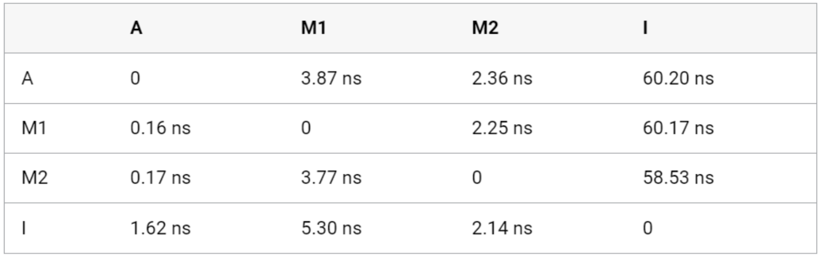


Table 1. MFPTs between Each Pair of the States in the A, M1, M2, and I States



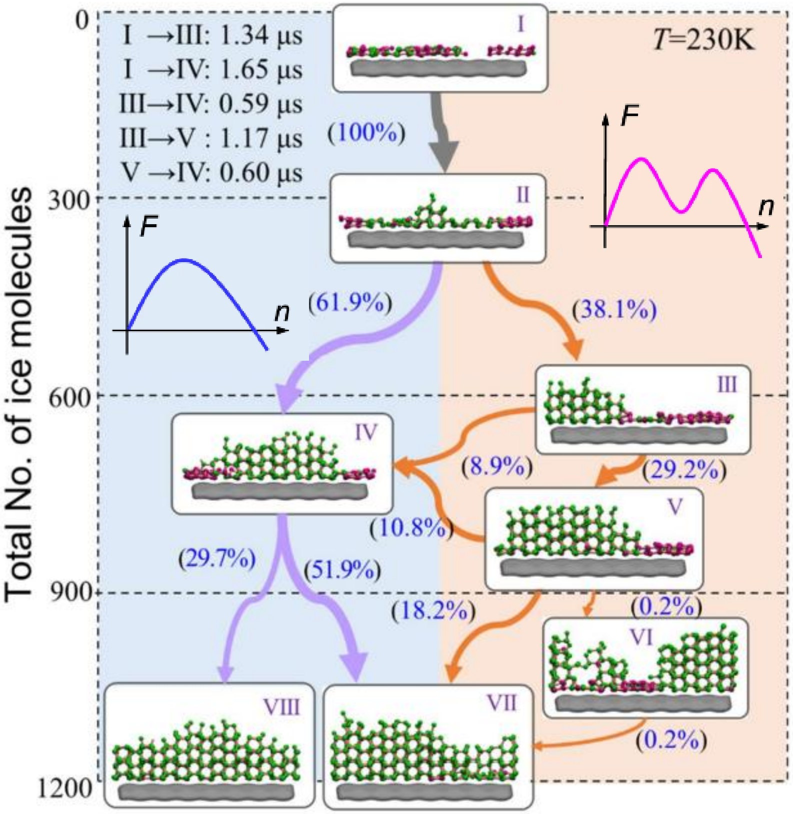
	A	M1	M2	I
A	0	3.87 ns	2.36 ns	60.20 ns
M1	0.16 ns	0	2.25 ns	60.17 ns
M2	0.17 ns	3.77 ns	0	58.53 ns
I	1.62 ns	5.30 ns	2.14 ns	0

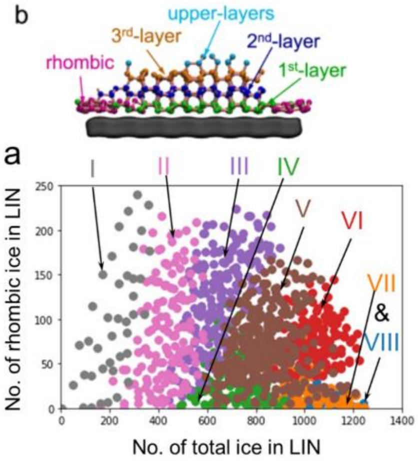
[1] J. Phys. Chem. B **2017**, 121, 14, 2952–2960

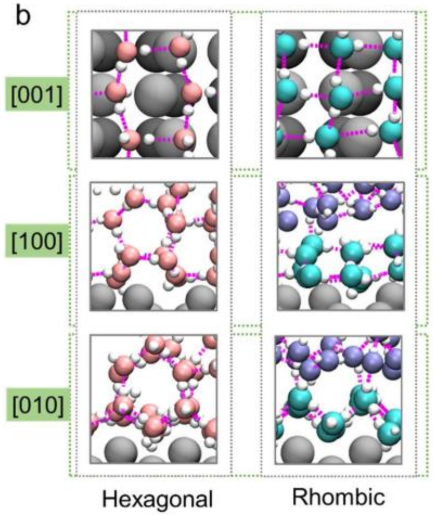
Analysis MSM : Let's watch the fancy applications

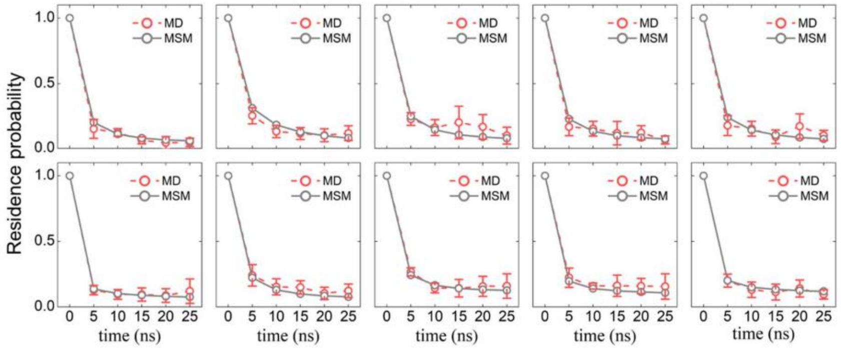
■ Let's see some references and check "What they calculated."

Paper 2 : Temperature-dependent kinetic pathways of heterogeneous ice nucleation competing between classical and non-classical nucleation









[1] *Nat Commun* **12**, 4954 (2021)

Content

Introduction) Markov chain : Memoryless

Building MSM : How do we partition the space and time?

Analysis MSM : What quantities can be calculated?

Further MSM : Wake up! It's time for math.

Further MSM : We have so many things to do! (with MATH)

■ Well, It’s time to response some questions. I bring 3 questions.

Non-Markovian Process : What if we allow the “memory”?

-> Generalized Master Equation (GME)

Tasting the Transition Probability Matrix : What is the other eigenvalues / eigenvectors, not its eigenvalue = 1?

Transition Path Theory : First time to meet “Committor”

[The lecture notes will be given.]

Non-Markovian Process : What if we allow the "memory"?

-> Generalized Master Equation (GME)

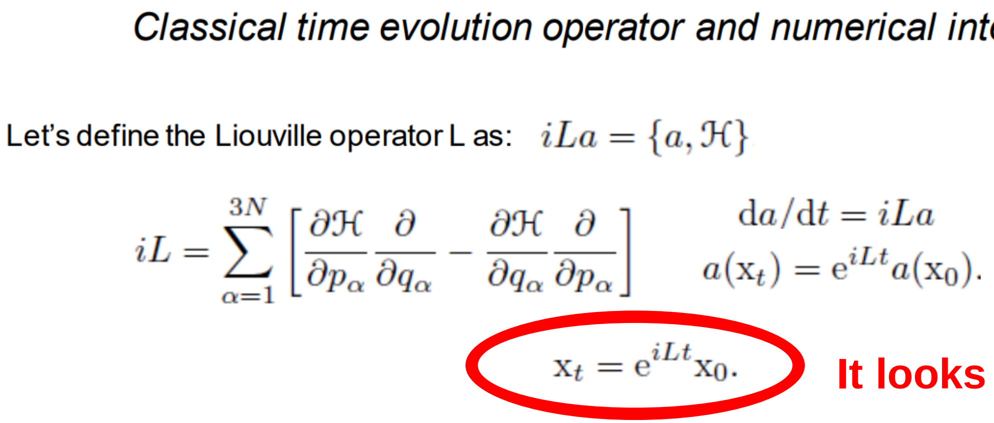
■ To construct TPM, we use lagged time. But, to satisfy the memoryless property, the lagged time should be longer than the relaxation time.

■ Some cases, running longer simulation than relaxation time is limited.

■ Then, What if we use shorter simulation and accept the memory property?

■ Wait, Can we ensure that the dynamics should be memoryless? Why?

Let's review SM's presentation on WEEK1 – **Liouville equation**



Classical time evolution operator and numerical integrators

Let's define the Liouville operator L as: $iLa = \{a, \mathcal{H}\}$

$$iL = \sum_{\alpha=1}^{3N} \left[\frac{\partial \mathcal{H}}{\partial p_\alpha} \frac{\partial}{\partial q_\alpha} - \frac{\partial \mathcal{H}}{\partial q_\alpha} \frac{\partial}{\partial p_\alpha} \right] \qquad \begin{aligned} \mathrm{d}a/\mathrm{d}t &= iLa \\ a(\mathrm{x}_t) &= \mathrm{e}^{iLt} a(\mathrm{x}_0). \end{aligned}$$

$$\mathrm{x}_t = \mathrm{e}^{iLt} \mathrm{x}_0.$$

It looks very similar with Markov chain!

[1] *J. Chem. Phys.* **2020**, 153, 014105.

Non-Markovian Process : What if we allow the "memory"?

-> Generalized Master Equation (GME)

■ Liouville's equation : $\frac{\partial \rho(t,\mathbf{\Gamma})}{\partial t} = \mathcal{L}\rho(t,\mathbf{\Gamma})$

$\rho(t,\Gamma)$: the probability distribution function across the entire phase space Γ at time t

■ The formula above means that $\rho(t+\tau,\mathbf{\Gamma}) = e^{\mathcal{L}\tau}\rho(t,\mathbf{\Gamma})$ → MEMORYLESS

Conclusion : In full dimension, the ensemble dynamics has memoryless property.

→ Can we project the Liouville operator to C.V.-space to make generalized master equation?

→ Hummer-Szabo projection operator :

$$\mathbb{P} = \sum_i |\chi_i(\boldsymbol{x})\rho_{eq}(\boldsymbol{x})\rangle \pi_i^{-1} \langle \chi_i(\boldsymbol{x})|$$

→ Nakajima-Zwanzig equation :

$$\frac{\partial}{\partial t}\mathbb{P}\rho(t) = \mathbb{P}\mathcal{L}\mathbb{P}\rho(t) + \mathbb{P}\mathcal{L}e^{\mathbb{Q}\mathcal{L}t}\mathbb{Q}\rho(0) + \int_0^t \mathbb{P}\mathcal{L}\, e^{\mathbb{Q}\mathcal{L}(t-s)}\mathbb{Q}\mathcal{L}\mathbb{P}\rho(s)\mathrm{ds}$$

→ General Master Equation(GME) : $\dot{\boldsymbol{T}}(t) = \boldsymbol{T}(t)\dot{\boldsymbol{T}}(0) - \int_0^t \boldsymbol{T}(t-\tau)\boldsymbol{K}(\tau)d\tau$

$$T_{ij}(t) = \langle \chi_j(\boldsymbol{x})|e^{\mathcal{L}t}|\chi_i(\boldsymbol{x})\rho_{eq}(\boldsymbol{x})\rangle \pi_i^{-1}$$

$$K_{ij}(t) = -\langle \chi_j(\boldsymbol{x})|\mathcal{L}e^{\mathbb{Q}\mathcal{L}t}\mathbb{Q}\mathcal{L}|\chi_i(\boldsymbol{x})\rho_{eq}(\boldsymbol{x})\rangle \pi_i^{-1}$$

[1] *J. Chem. Phys.* **2020**, 153, 014105.

Tasting the Transition Probability Matrix : What is the other eigenvalues / eigenvectors, not its eigenvalue = 1?

■ TPM, # of states = n, has n, nondegenerate Left and right eigenvectors, whose eigenvalues are $|\lambda| \leq 1$

■ Eigenvalues are related to the relaxation time of each state.

$\boldsymbol{r}_m$ $(m = 1, \ldots, n\ )$: right eigenvectors of $\boldsymbol{T}$ → eigenvalues : $\lambda_1, \ldots, \lambda_n$
$\boldsymbol{r}_m$: orthonormal basis w.r.t. the weighted inner product

$$\begin{aligned} \mathbf{T}\mathbf{r}_m &= \lambda_m \mathbf{r}_m, \\ \langle \mathbf{r}_m, \mathbf{r}_{m'} \rangle_\pi &= \delta_{m,m'}. \end{aligned}$$

* weighted inner product

$$\langle \mathbf{v}, \mathbf{w} \rangle_\pi = \sum_{i=1}^{n} v_i w_i \pi_i,$$

Then, left eigenvectors $\boldsymbol{l}_m \coloneqq \mathbf{\Pi}\mathbf{r}_\mathrm{m}$ exists and using spectral decomposition,

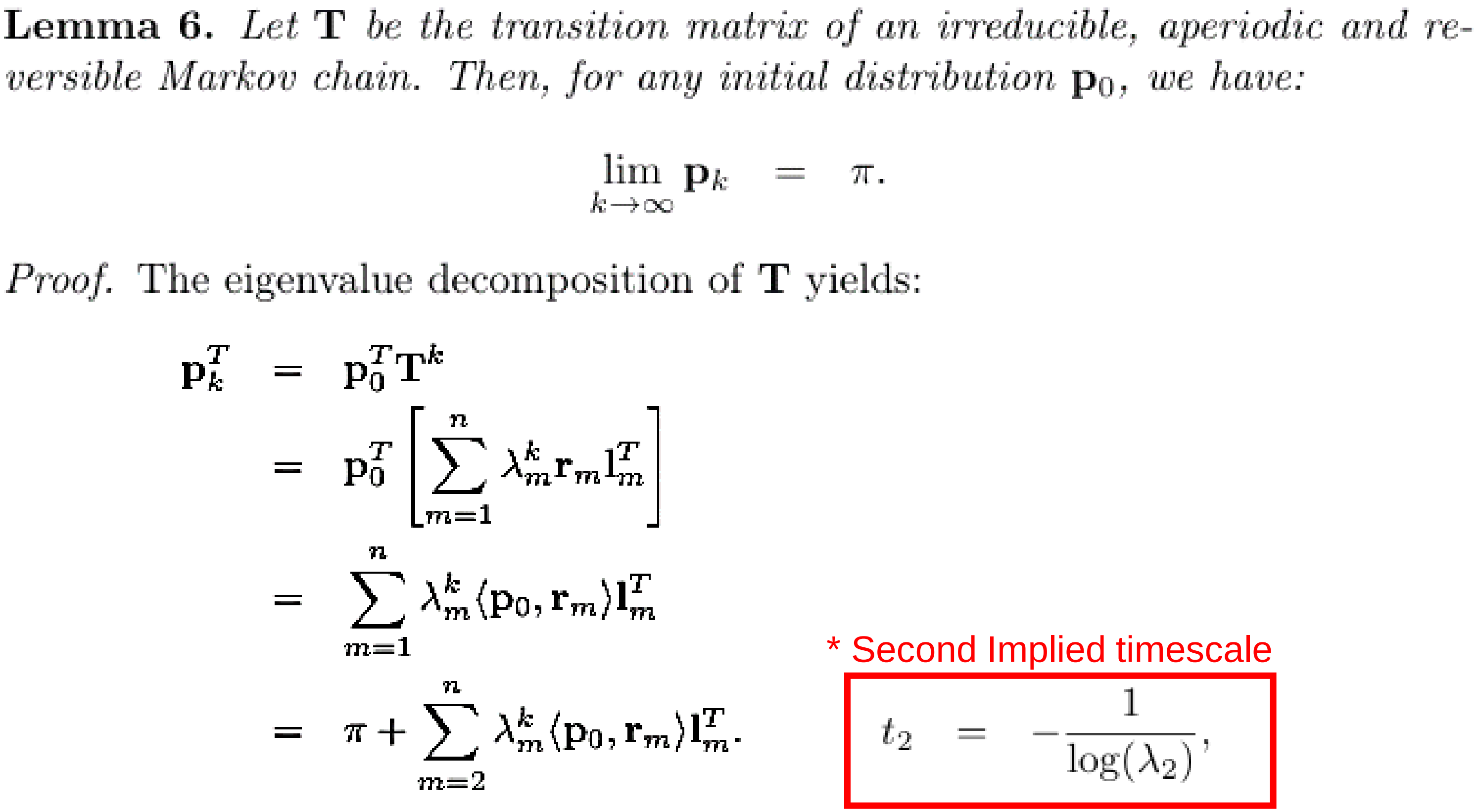
$$\begin{aligned} \mathbf{T}(i,j) &= \sum_{m=1}^{n} \lambda_m \mathbf{r}_m(i) \pi(j) \mathbf{r}_m(j) \\ &= \sum_{m=1}^{n} \lambda_m \mathbf{r}_m(i) \mathbf{l}_m(j). \\ \mathbf{T} &= \sum_{m=1}^{n} \lambda_m \mathbf{r}_m \mathbf{l}_m^T. \end{aligned}$$

→ NEXT

[1] *J. Chem. Phys.*, **2011**, 134, 174105.

Tasting the Transition Probability Matrix : What is the other eigenvalues / eigenvectors, not its eigenvalue = 1?

■ Using the spectral decomposition of T matrix, We can derive the convergence of any initial state p_0 to state π



Lemma 6. *Let* $\mathbf{T}$ *be the transition matrix of an irreducible, aperiodic and reversible Markov chain. Then, for any initial distribution* $\mathbf{p}_0$*, we have:*

$$\lim_{k\to\infty} \mathbf{p}_k = \pi.$$

Proof. The eigenvalue decomposition of $\mathbf{T}$ yields:

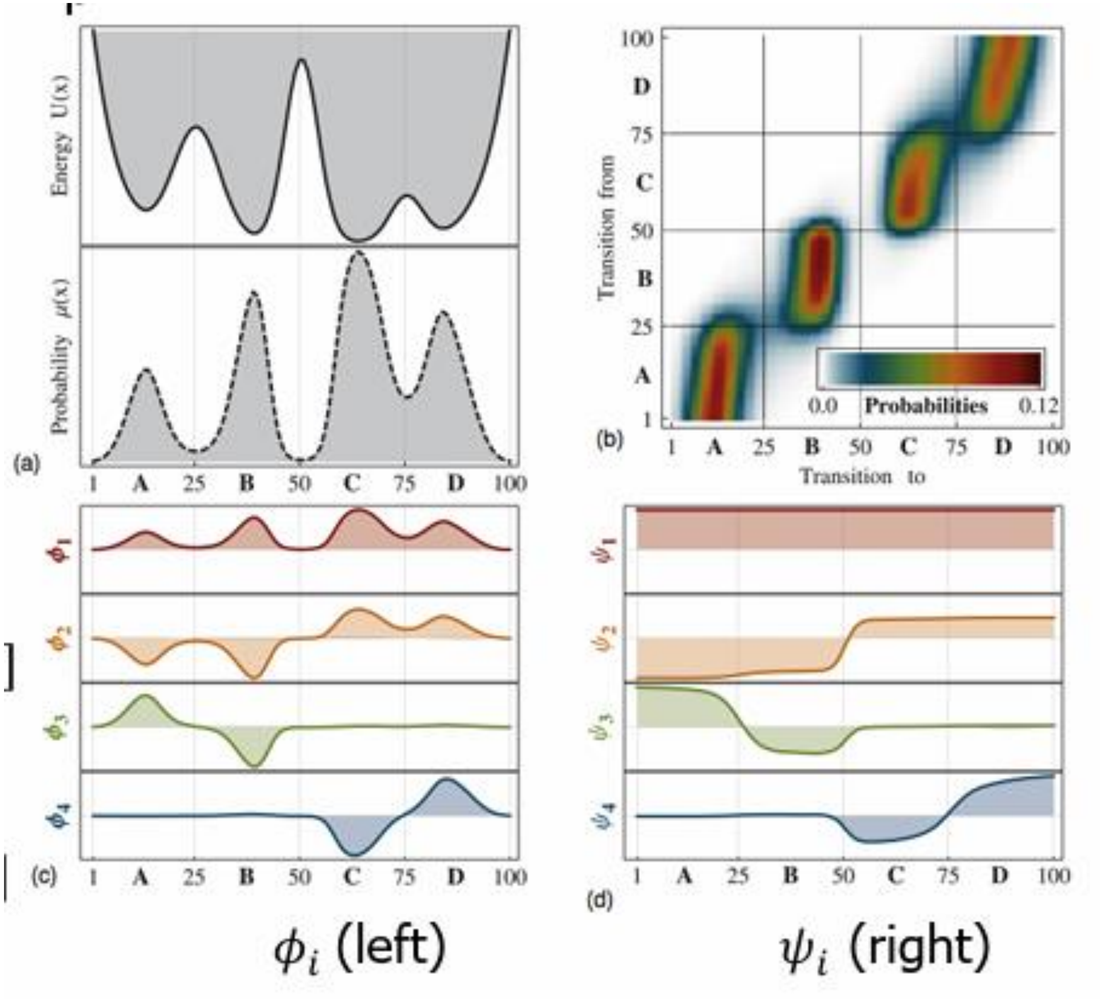
$$\begin{aligned}
\mathbf{p}_k^T &= \mathbf{p}_0^T \mathbf{T}^k \\
&= \mathbf{p}_0^T \left[\sum_{m=1}^{n} \lambda_m^k \mathbf{r}_m \mathbf{l}_m^T \right] \\
&= \sum_{m=1}^{n} \lambda_m^k \langle \mathbf{p}_0, \mathbf{r}_m \rangle \mathbf{l}_m^T \\
&= \pi + \sum_{m=2}^{n} \lambda_m^k \langle \mathbf{p}_0, \mathbf{r}_m \rangle \mathbf{l}_m^T.
\end{aligned}$$

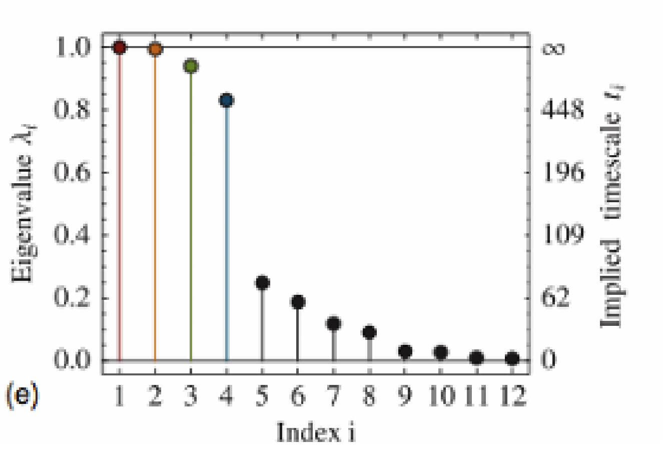
* Second Implied timescale

$$\boxed{t_2 = -\frac{1}{\log(\lambda_2)},}$$

[1] *J. Chem. Phys.*, **2011**, 134, 174105.

Tasting the Transition Probability Matrix : What is the other eigenvalues / eigenvectors, not its eigenvalue = 1?





Eigenfunction points to the location of metastable states

[1] *J. Chem. Phys.*, **2011**, 134, 174105.

Transition Path Theory : First time to meet "Committor"

- Transition Path Theory (TPT) : Find the paths between state A and B!

- Hitting Probabilities and Committors

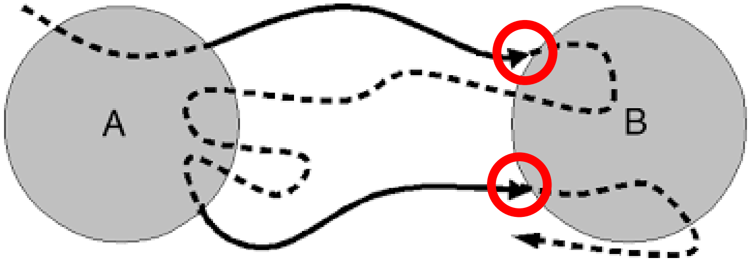
$$H_A = \min\{k \geq 0 : X_k \in A\},$$: Hitting time of a set A

$$h_A(i) = \mathbb{P}_i(H_A(i) < \infty).$$: the corresponding hitting probability which starts at state i

- Forward committor : the probability to hit set B next rather than A.

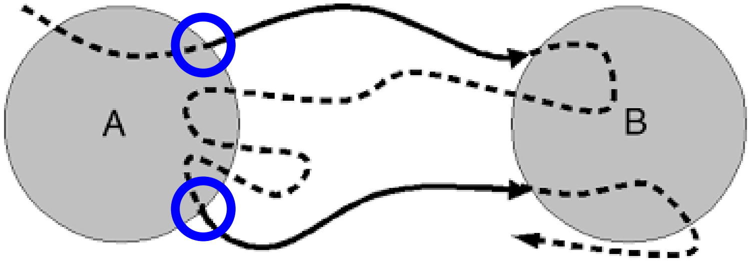
$$q_i^+ = \mathbb{P}_i(H_B < H_A).$$

$$q_i^+ = \begin{cases} 1, & i \in B \\ 0, & i \in A \\ \sum_{j=1}^n \mathrm{T}_{ij} q_j^+, & \textit{otherwise.} \end{cases}$$



- Backward committor : the probability to come from A rather than from B.

$$q_i^- = \begin{cases} 0, & i \in B \\ 1, & i \in A \\ \sum_{j=1}^n \mathrm{T}_{ij} q_j^-, & \textit{otherwise.} \end{cases}$$



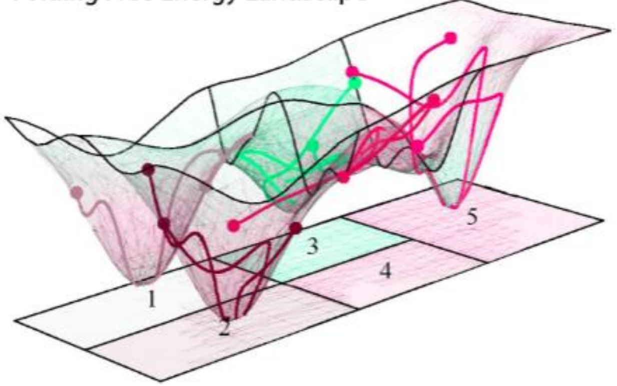
[1] Multiscale Model. Simul. **2009**, 7, 1192– 1219

Transition Path Theory : First time to meet "Committor"

- Fluxes and Transition Rates can be calculated by forward/backward committors.
- Probability current between states I and J $f_{ij}^{AB} = \begin{cases} \pi_i q_i^- \mathrm{T}_{ij} q_j^+, & i \neq j \\ 0, & \text{otherwise.} \end{cases}$

(Effective probability current : $f_{ij}^{+} = \max\left(f_{ij}^{AB} - f_{ji}^{AB}, 0\right)$)

- Average total number of trajectories: $F^{AB} = \sum_{i \in A} \sum_{j \in S} f_{ij}^{AB}$
- transition rate: $\kappa_{AB} = \frac{F^{AB}}{\sum_{j \in S} \pi_j q_j^-}.$
- MFPT : the inverse of the transition rate : $\tau_{AB} = \kappa_{AB}^{-1}$



Folding Free Energy Landscape

■ Finding dominant pathways:

Let) $w = (i_0, i_1, \dots, i_K)$: simple reaction pathway ($i_0 \in A, i_K \in B, i_1, \dots, i_{K-1} \in (A \cup B)^c$)

■ Min-current (=capacity): $\boxed{c(w) = \min_{(i,j) \in w} f_{ij}^{+},}$

➔ Edge (i, j) where minimum current occurs : **bottleneck**
➔ Bast pathway : one which maximizes the min-current

[1] Multiscale Model. Simul. **2009**, 7, 1192– 1219

Takeaways

The most important property of markovian process is “MEMORYLESS”.

After we set the TPM of MSMs, we can calculate many thermodynamic & kinetic quantities.

Unlike our expectations, It is difficult to satisfy memoryless property.

To describe non-markovian process, we can use other sophisticated methods, like GME.

Transition Path Theory (TPT) offers the theoretical frame : How to interpret

By using TPT, we can calculate MFPT, best pathway, .. Etc.

Q&A