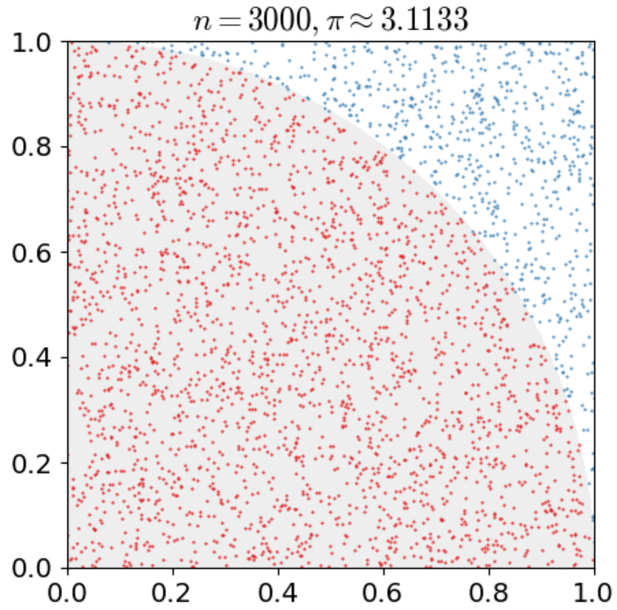
1/12/2021 Week 1 Module 2
International Workshop on Modeling Biological Macromolecules

- This module will consist of a
  - mini-lecture on molecular simulation including Markov Chain Monte Carlo, Molecular Dynamics and Hybrid Monte Carlo
  - walk-through of a python script to run HMC with Robosample
- At the end of this module, you should be able to address these questions:
  - What is Markov Chain Monte Carlo and Molecular Dynamics?
  - What is Hybrid Monte Carlo and why do people use it?
  - Generally speaking, how does a Hybrid Monte Carlo simulation work?
- You should also be able to run a simulation of a simple system using Robosample

Monte Carlo



- Monte Carlo simulation:
  - named after famous gambling city
  - uses random numbers
  - usually applied to hard deterministic or probabilistic problems
- Examples:
  - pi approximation
  - virtually tossing a coin or rolling a dice many times
  - estimating financial risk (uncertainty in unit price, sales…)
  - solving integrals / differential equations
- Cannot use it for complex highly dimensional probability distribution

Markov Chains

- Stochastic process: Sequence of random variables mapped to another variable (t)

  $\boldsymbol{X}$ = stochastic proccess; $\boldsymbol{\chi}$ = state space; $\boldsymbol{\pi}$ = probability vector; $\boldsymbol{\pi} = [\pi(A), \pi(B)]$, $\pi(A) = P(X = A)$

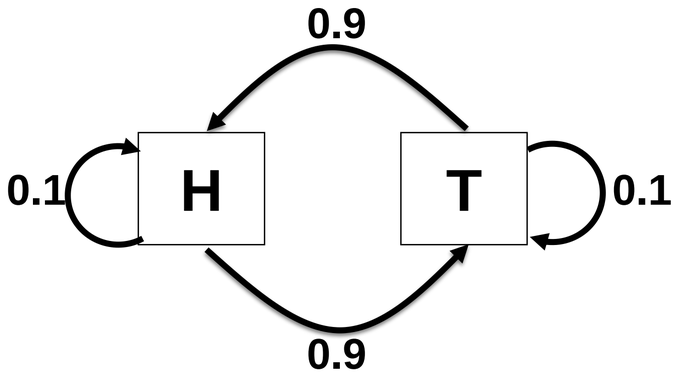
  $\boldsymbol{X} = \{X_0, X_1, X_2 \ldots.\}$, $X_i = x \in \boldsymbol{\chi}$, $\boldsymbol{\pi}\,(\boldsymbol{\chi})$

- Markovian property: memoryless: future only depends on the present

  $P(X_{i+1}|X_i = x_i, X_{i-1} = x_{i-1} \ldots, X_0 = x_0) = P(X_{i+1}|X_i = x_i)$

- Chain diagrams and transition matrix

diagram

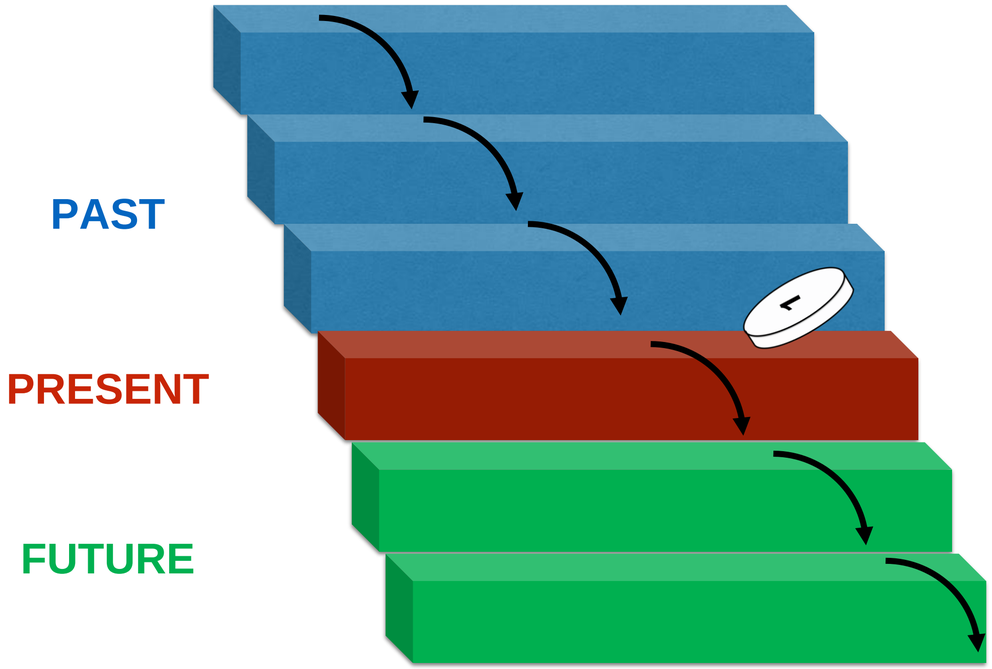


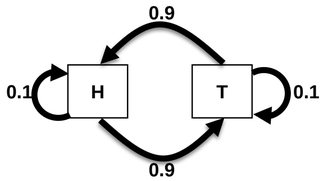
transition matrix

$$Q = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$

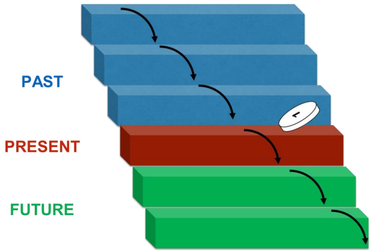
probability vector evolution

$$\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k Q$$



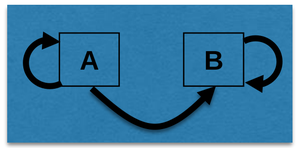


Markov Chains



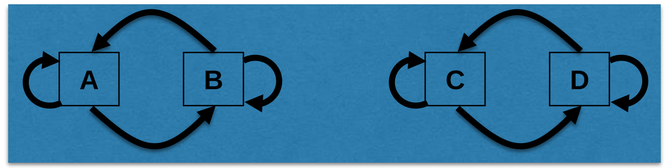
- Stationary probability vector (equilibrium): $\boldsymbol{\pi} = [\pi(A), \pi(B)]$, $\pi(A) = P(X = A)$

  $\boldsymbol{\pi} Q = \boldsymbol{\pi}$

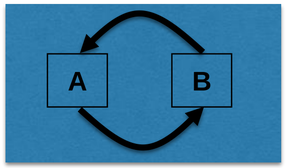


- Homogenous: transition probabilities are constant

- Irreducibility (weakly connected with no absorbing states)



- Periodicity



- Recurrent / transient state (always / never come back). If all states are recurrent the chain is ergodic.

- Detailed balance guarantees convergence to a stationarity distribution

$$\pi(A)P(A \to B) = \pi(B)P(B \to A)$$

Metropolis – Hastings Algorithm Derivation

- Markov Chain Monte Carlo usefulness: construct our own Markov Chain that converges to a specific probability distribution.

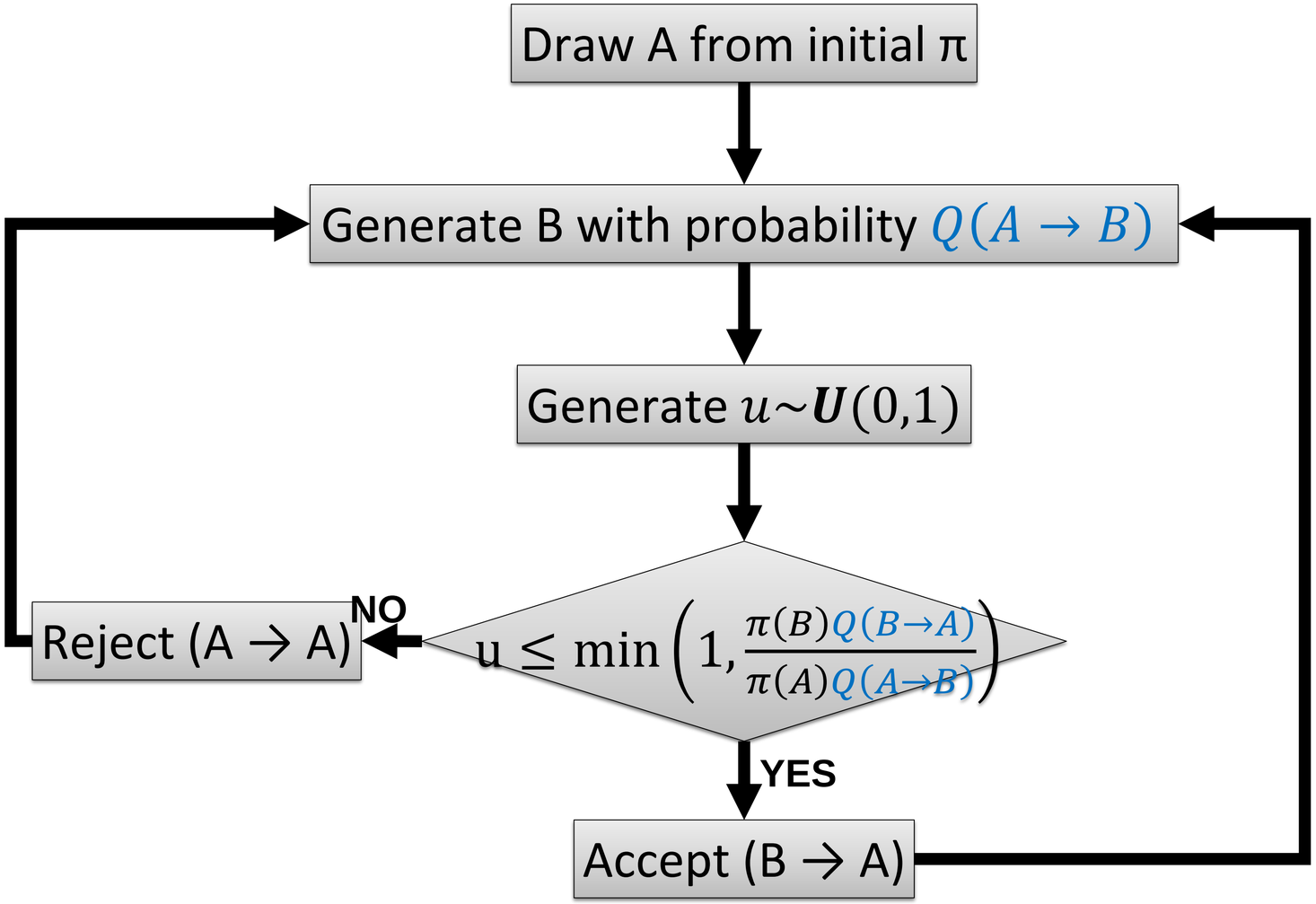
detailed balance $\pi(A)P(A \to B) = \pi(B)P(B \to A)$

proposal distribution $\pi(A)Q(A \to B) \neq \pi(B)Q(B \to A)$

correcting term $\pi(A)Q(A \to B)\alpha(A,B) = \pi(B)Q(B \to A)\alpha(B,A) \Leftrightarrow \frac{\alpha(A,B)}{\alpha(B,A)} = \frac{\pi(B)Q(B \to A)}{\pi(A)Q(A \to B)}$

Metropolis-Hastings $\alpha(A,B) = \min\left(1, \frac{\pi(B)Q(B \to A)}{\pi(A)Q(A \to B)}\right)$

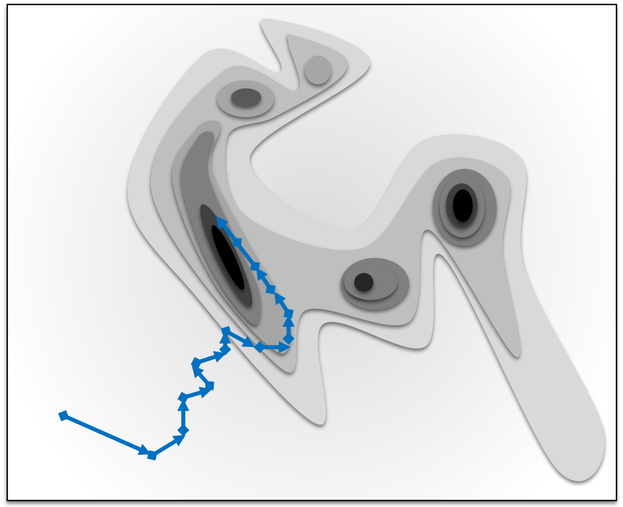
Metropolis – Hastings Algorithm



MCMC In Practice:
When do we start recording?

- The initial distribution is not known. Burn-in
  - achieve stationarity within a certain threshold
  - get to a high probability region

$$\bar{X}_n \to \mu \text{ as } n \to \infty$$



MCMC In Practice: When do we stop recording?

- Convergence of π^N towards the limiting distribution π_{true} $\quad dist(\pi^N, \pi_{true}) \leq \varepsilon$
  - Total variation distance (TVD)

$$dist(\pi^N, \pi_{true}) = \frac{1}{2}\sum_{\chi}|\pi^N(x) - \pi_{true}(x)| \qquad dist(\pi^N, \pi_{true}) = \sup(\pi^N(E) - \pi_{true}(E)),$$
E is any event

  - Hellinger distance

$$dist(\pi^N, \pi_{true}) = \frac{1}{\sqrt{2}}\sqrt{\sum_{\chi}\left(\sqrt{\pi^N(x)} - \sqrt{\pi_{true}(x)}\right)^2}$$

  - Relative entropy (information gain, Kullback–Leibler divergence)

$$dist(\pi^N, \pi_{true}) = \sum_{x}\pi^N(x)\,[-\log\pi_{true}(x)] - \left(\sum_{x}\pi^N(x)[-\log\pi^N(x)]\right)$$

  - Mixing time

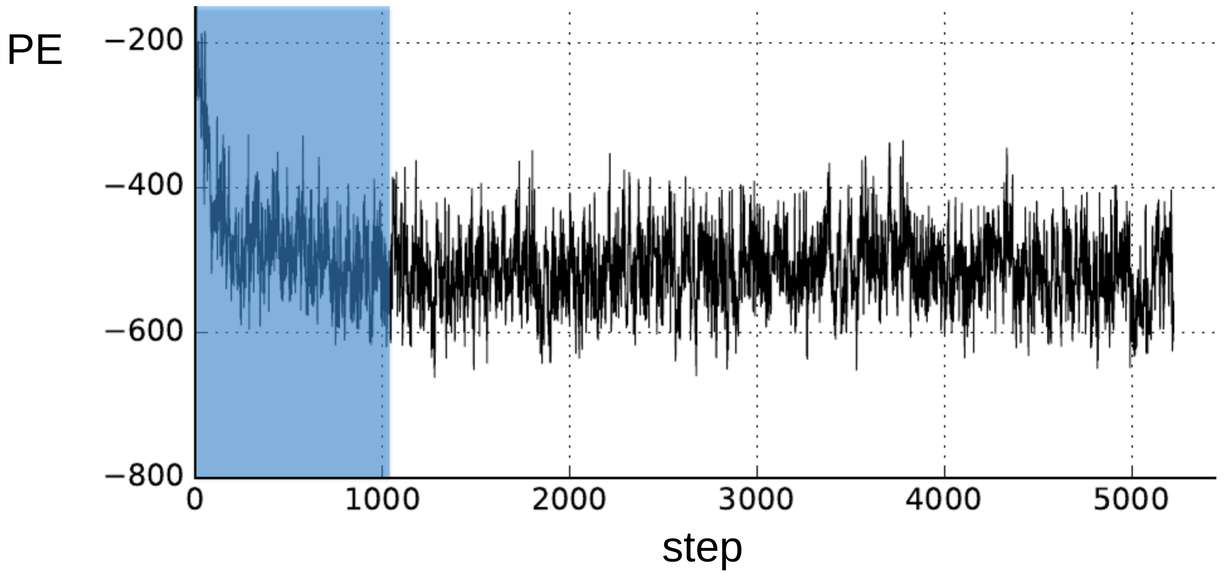
$$t_{mix}(\varepsilon) = \min\{t: dist(t) \leq \epsilon\}$$

Levin Peres and Wilmer, “Markov Chains and Mixing Times”

MCMC Simulation In Practice. When do we stop recording?

- Convergence of observables tests when the reference π_{true} is not known

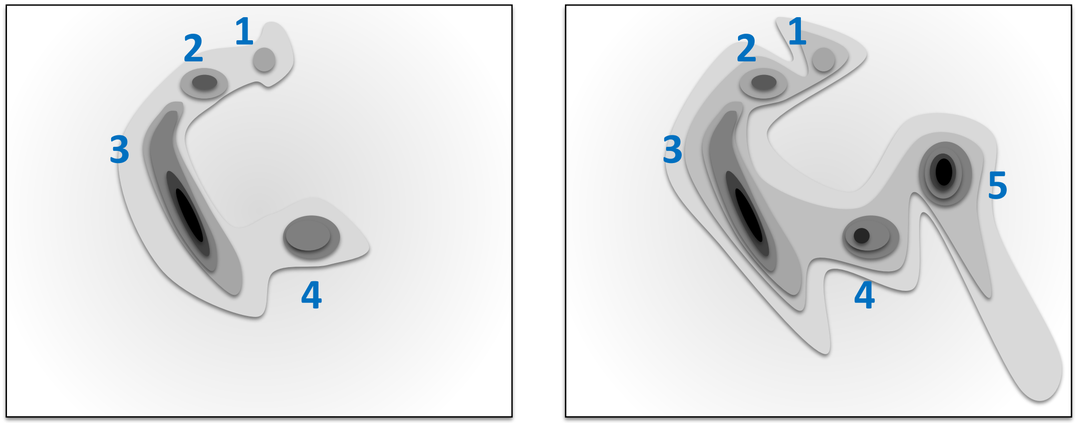
$$\mathcal{O} \rightarrow O_{ct}$$



Sawle and Ghosh, "Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma."
Grossfield A, Zuckerman DM. Quantifying uncertainty and sampling quality in biomolecular simulations.

MCMC Simulation In Practice. When do we stop recording?

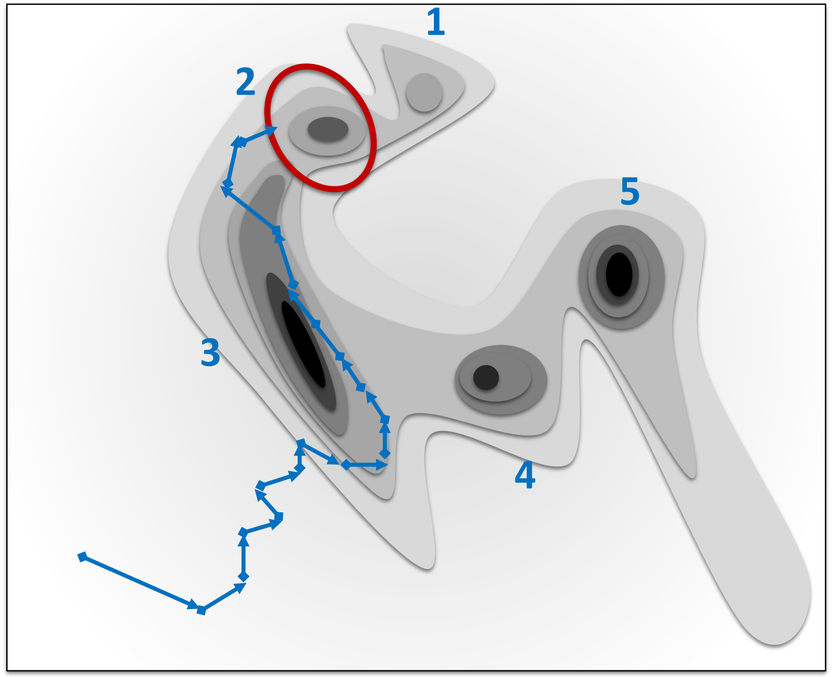
- Self-consistency tests: “monitoring the overlap between full and partial trajectories”. E.g. constant number of clusters.



Sawle and Ghosh, “Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma.”
Grossfield A, Zuckerman DM. Quantifying uncertainty and sampling quality in biomolecular simulations.

MCMC In Practice: How efficient is it?

- Correlation time analysis: time required to lose memory of previous values
- Hitting time
- Cover time
- Mean first passage matrix



	1	2	3	4	5
1					
2					
3					
4					
5					

What is Molecular Dynamics?

- Add energy to a system modeled by molecular mechanics and simulate its progress with time using Newton's second law of motion $\vec{F} = ma$
- See 0:45 to 2:20 of "An Introduction to Molecular Dynamics" (https://www.youtube.com/watch?v=ILFEqKl3sm4)
- See a separation of alkane and water: https://www.youtube.com/watch?v=xcMSHy3CqXA

Why do biological molecular dynamics?

- “everything that living things do can be understood in terms of the jigglings and wigglings of atoms” - Richard Feynman

- Check out David’s molecular dynamics YouTube playlist: https://www.youtube.com/playlist?list=PLYEyeVFrqfAu0ft6sF4vVe5FOEbJYyi_L

General MD Algorithm

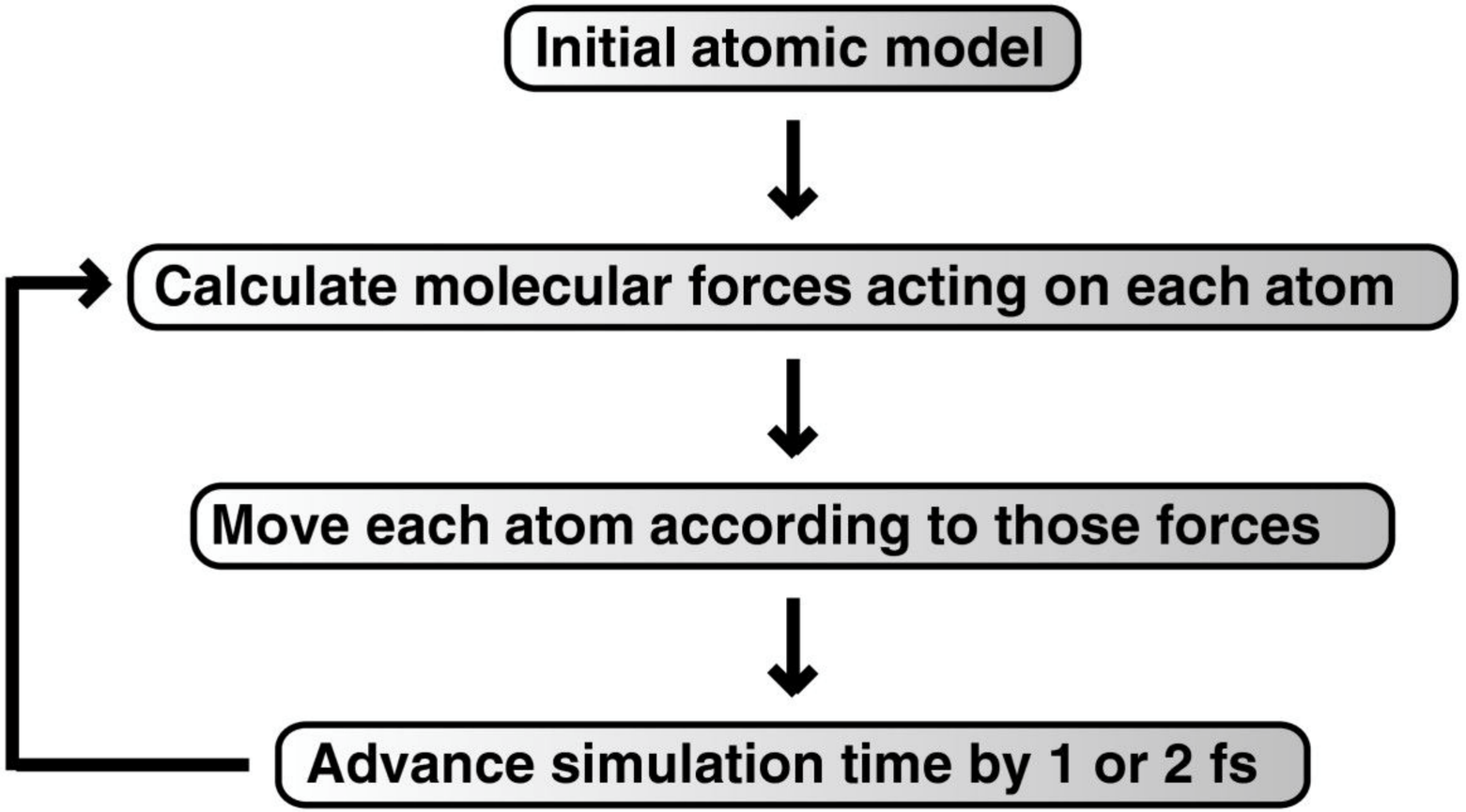


Figure 2 of Durrant & McCammon, 2011

How do we calculate trajectories?

- Evolution in time given is by classical mechanics: Hamilton's equations

$$\mathbb{T}:$$

$$Force: \quad \frac{dp}{dt} = -\frac{\partial \mathcal{H}}{\partial x}$$

$$Velocity: \quad \frac{dx}{dt} = \frac{\partial \mathcal{H}}{\partial p}$$

- Integrate trajectory using Taylor expansion

$$x(t) = \frac{1}{0!}x(t_0)(t-t_0)^0 + \frac{1}{1!}\frac{dx}{dt}(t_0)(t-t_0)^1 + \frac{1}{2!}\frac{d^2x}{dt^2}(t_0)(t-t_0)^2 + \frac{1}{3!}\frac{d^3x}{dt^3}(t_0)(t-t_0)^3 + \cdots$$

Ergodic Hypothesis

- Evolution in time given by classical mechanics: Hamilton's equations

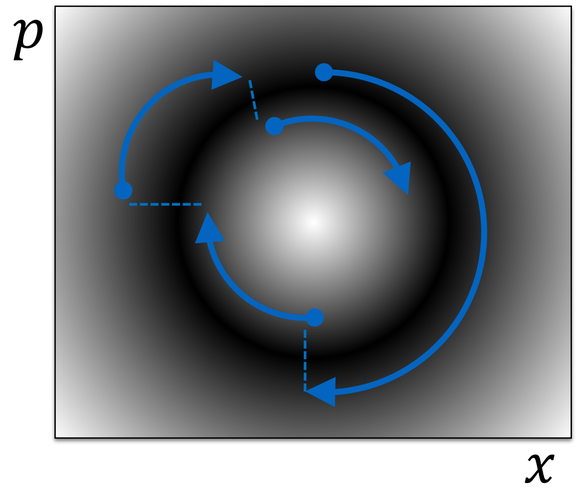
$$\mathbb{T}:$$

$$Force: \quad \frac{dp}{dt} = -\frac{\partial \mathcal{H}}{\partial r}$$

$$Velocity: \quad \frac{dr}{dt} = \frac{\partial \mathcal{H}}{\partial p}$$

- Time averages equals space averages

$$\frac{1}{t}\int_0^t \mathcal{O}(\mathbb{T}^s(r_0, p_0))ds = \int_\Gamma \pi(r,p)\mathcal{O}(r,p)drdp$$



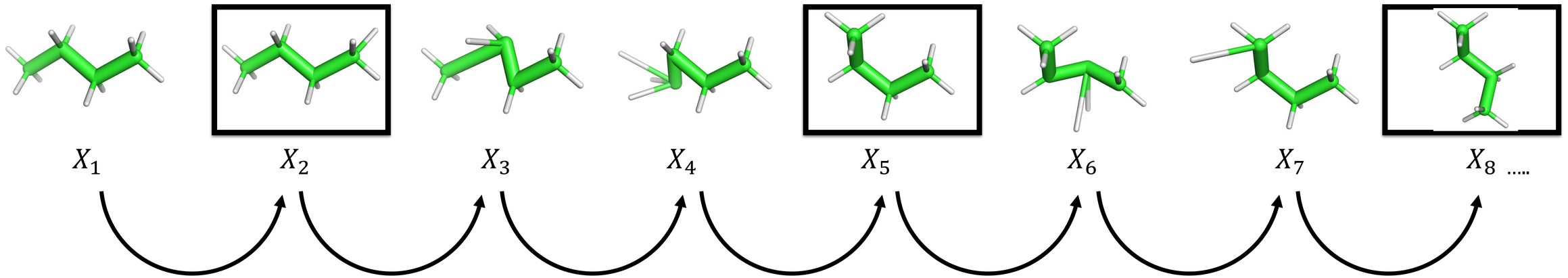
Disscussion

- MD:
  - - correct model of molecular behavior
  - - only approximates the desired probability distribution

- MCMC:
- - doesn’t provide behavior of molecules
- - guarantees samples from the desired distribution

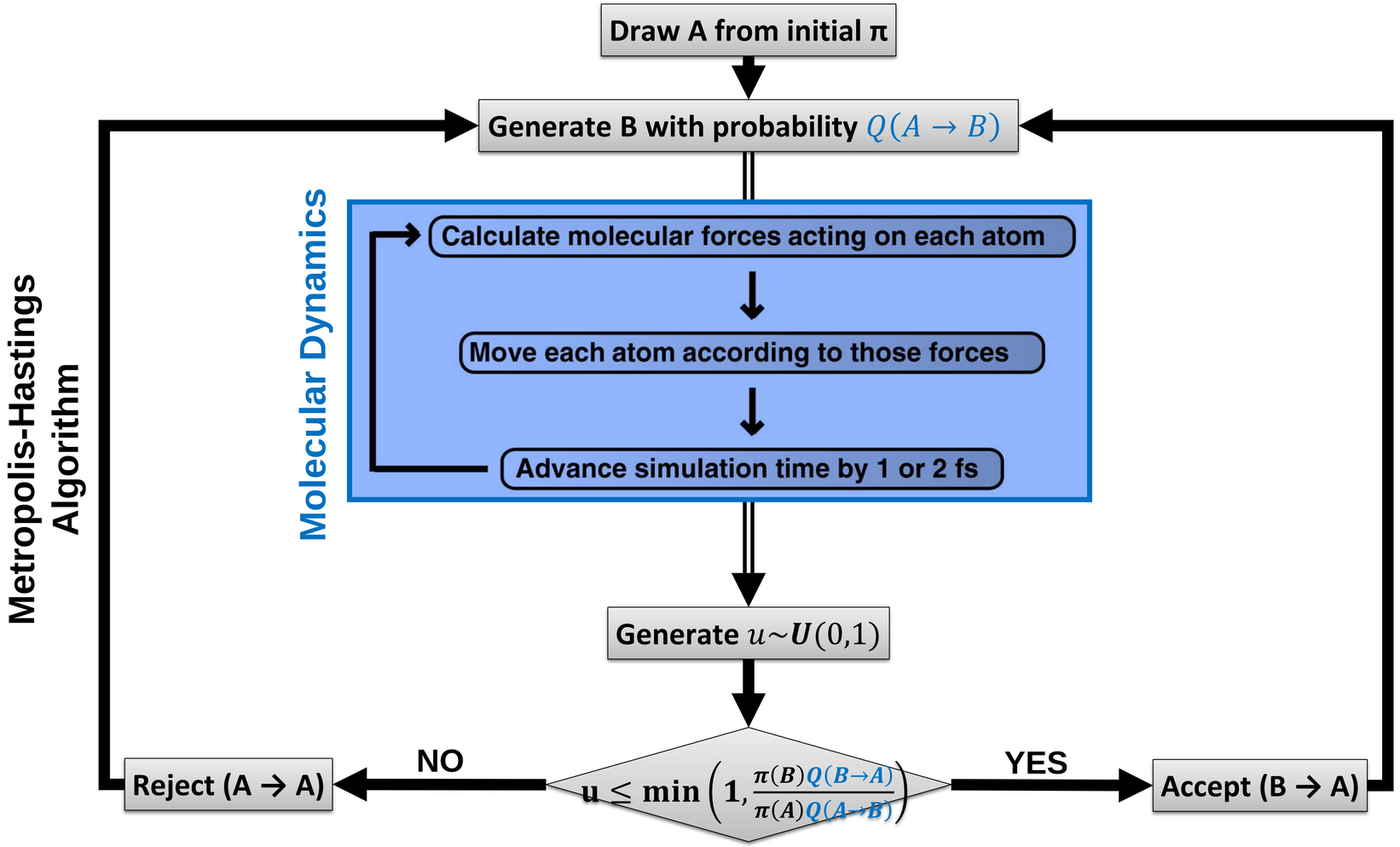
Markov Chain Monte Carlo for biological molecules

- Randomly chosen configurations lead to low acceptance rates



- Can we do better?

Hybrid Monte Carlo



Why does it work?

- Sample from the joint distribution $\pi(\boldsymbol{r}, \boldsymbol{p})$ and use the marginal $\pi(\boldsymbol{r})$ because $\boldsymbol{r}$ does not depend on p

$$\pi(\boldsymbol{r}, p) \propto e^{-\beta U(\boldsymbol{r})} e^{-\beta \frac{1}{2} p^T M^{-1} p} = e^{-\beta U(\boldsymbol{r})} \mathcal{N}(0, M)$$

- What potential energy function should we use?
  - solve for potential energy above and get $\mathrm{U}(\boldsymbol{r}) = -\log[\pi(\boldsymbol{r})]$

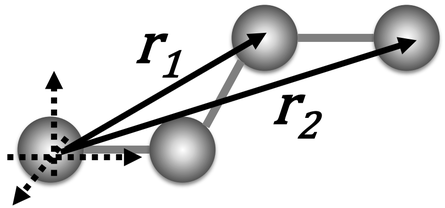
Why use constraints?

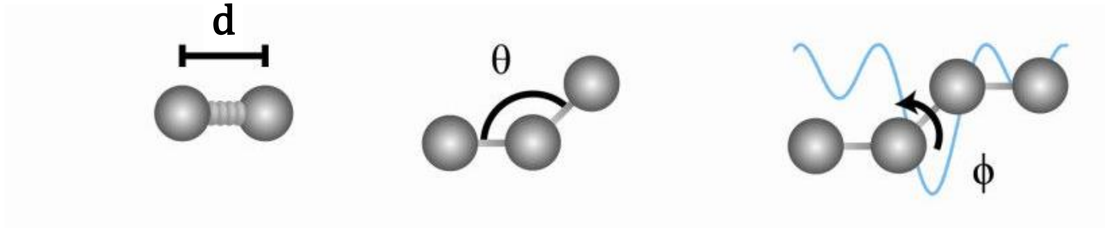
- Target distribution is highly dimensional and too complex to get conclusive results in reasonable amount of time
- Heavier bodies allows the increase of the timestep
- E.g. :
  - rigid water molecules (TIP3P model)
  - constant bond lengths and bond angles: torsional dynamics
  - constrain specific regions of molecules or even entire domains

Taken from Figure 3 of Durrant & McCammon, 2011.

How to impose constraints?

- Cartesian coordinates and internal coordinates





- Dynamics with maximal coordinates (Lagrange multipliers)

$$\mathbb{T}:$$

$$Force: \quad \frac{dp}{dt} = -\frac{\partial \mathcal{H} + \lambda c(r)}{\partial r} \qquad c(r) = \|r_1 - r_2\|^2 - d^2$$

$$Velocity: \frac{dr}{dt} = \frac{\partial \mathcal{H}}{\partial p}$$

$$r(t) = \frac{1}{0!} r(t_0)(t - t_0)^0 + \frac{1}{1!}\frac{dr}{dt}(t_0)(t - t_0)^1 + \frac{1}{2!}\frac{d^2 r}{dt^2}(t_0)(t - t_0)^2 + \cdots$$

- Example SHAKE algorithm usually used for hydrogen atoms bonds

Taken from Figure 3 of Durrant & McCammon, 2011.

How to impose constraints?

- Dynamics reduced coordinates (Featherstone)

$$\mathbb{T}:$$

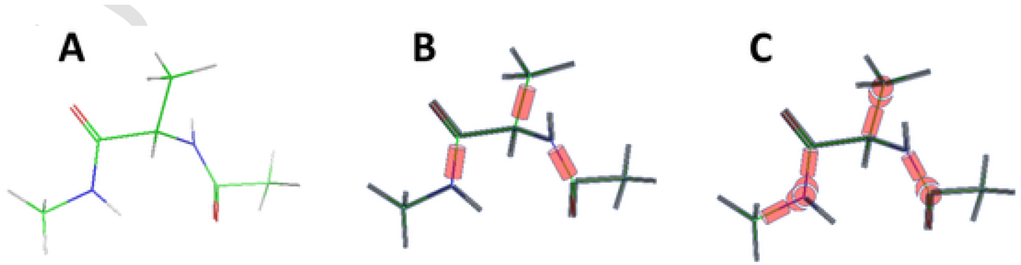
$$Force: \quad \frac{dp}{dt} = -\frac{\partial \mathcal{H}}{\partial \phi}$$

$$Velocity: \quad \frac{d\phi}{dt} = \frac{\partial \mathcal{H}}{\partial p}$$

$$\phi(t) = \frac{1}{0!}\phi(t_0)(t-t_0)^0 + \frac{1}{1!}\frac{d\phi}{dt}(t_0)(t-t_0)^1 + \frac{1}{2!}\frac{d^2\phi}{dt^2}(t_0)(t-t_0)^2 + \cdots$$

- Rigid body dynamics includes rotational quantities which are incorporated using Euler's laws of motion

- Dynamics is altered



Taken from Figure 3 of Durrant & McCammon, 2011.

Gibbs sampling

- Why?
  - Only simulating with constraints is not enough. The simulation does not cover the entire conformational space
  - Sampling from complex multivariate joint probability.
- How?
  - Take turns in sampling from conditionals. Allow oversampling easier to sample variables.

$$1.\ \pi(X|Y)$$
$$2.\ \pi(Y|X)$$

- Robosample scheme: constrained dynamics combined with all-atom dynamics

$$1.\ \pi(\phi|d,\theta)$$
$$2.\ \pi(d,\theta,\phi)$$

Why Robosample?

- Many choices of software for molecular dynamics
  - https://en.wikipedia.org/wiki/Comparison_of_software_for_molecular_mechanics_modeling
  - https://www.rcsb.org/pages/thirdparty/modeling_and_simulation
- Robosample is
  - free
  - GPU-accelerated
  - can be used in python scripts/C++ programs

Review Questions

- Generally speaking, how does a HMC simulation work?

References

- Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. BMC Biol 2011, 9 (1), 71. https://doi.org/10.1186/1741-7007-9-71, adapted under the CC BY 2.0 license.
- Kalos, Malvin H.; Whitlock, Paula A. (2008). Monte Carlo Methods.
- Levin DA, Peres Y and Wilmer EL, “Markov Chains and Mixing Times”, American Mathematical Soc., Oct 31, 2017, ISBN: 1470429624, 9781470429621
- Sawle L, Ghosh K. Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma. J Chem Theory Comput. 2016 Feb 9;12(2):861-9. doi: 10.1021/acs.jctc.5b00999. Epub 2016 Jan 26. PMID: 26765584.
- Grossfield A, Zuckerman DM. Quantifying uncertainty and sampling quality in biomolecular simulations. Annu Rep Comput Chem. 2009 Jan 1;5:23-48. doi: 10.1016/S1574-1400(09)00502-7. PMID: 20454547; PMCID: PMC2865156.