Hierarchical Metastable States and Kinetic Transition Networks:

Trajectory Mapping and Clustering

Xin Zhou

xzhou@gucas.ac.cn

Graduate University of Chinese Academy of Sciences, Beijing

2012.6.5 KITP, Santa Barbara

中國科學院研究生院
GRADUATE UNIVERSITY OF CHINESE ACADEMY OF SCIENCES

Multiscale modeling and simulation:

- Understanding systems from simulation data (Date Mining)
- coarse-graining
- enhance sampling and accelerating dynamics

Understand MD Simulation data

- High dimension
- Large number of configurations
- Complicate structure

Project to low dimension: states and transitions

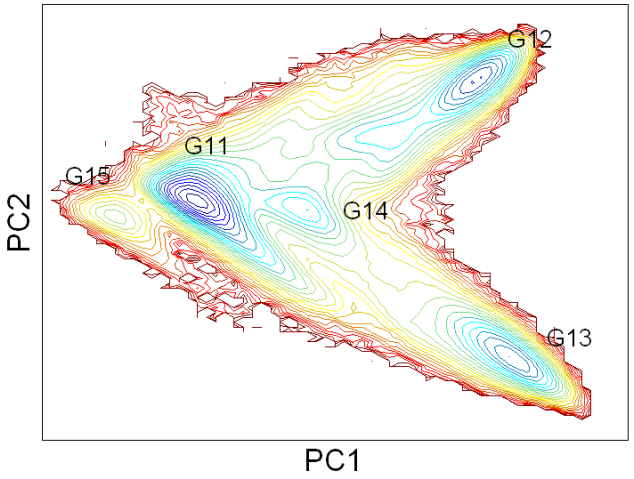
End-End distance,

Radius of gyration,

Root mean squared distance (RMSD)

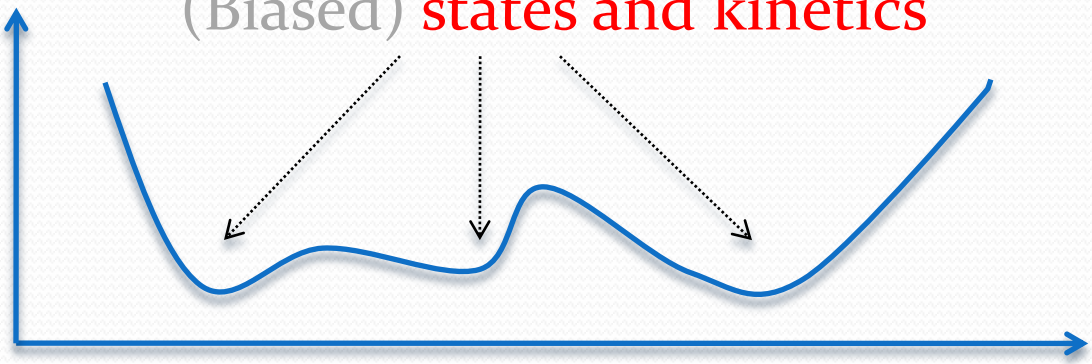
Principle Components (PC)

.....



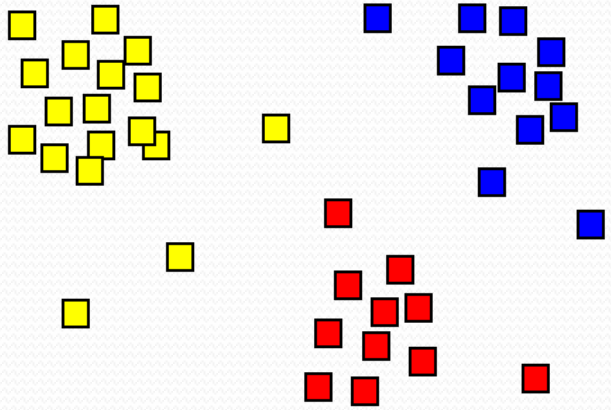
Reaction coordinates are usually hard to know a priori

(Biased) states and kinetics



Cluster Analysis

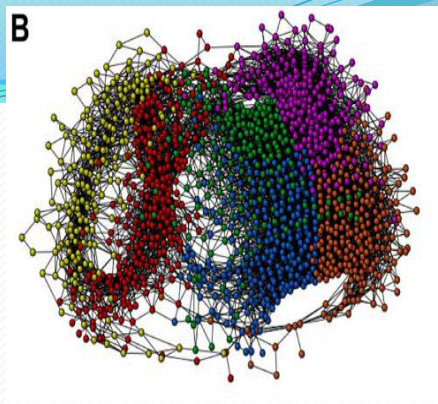
1. Get the distance matrix
2. clustering



Hundred kinds of clustering algorithms

Not very stable, have some adjusted parameters

Configuration Cluster Analysis (*Markov Chain Model)*



- Divide MD configurations into lots of microstates
- Estimate transition rate matrix among these microstates
- Clustering these microstates to a few macro-states

Bottom-up

F. Noe, S. Fischer, Cur. Opin. Struc. Biol. (2008);
D. Gfeller, P. DeLosRios, A. Caflisch, PNAS (2007);
D. Prada-Gracia et al, PLoS Comp. Biol. (2009);
F. Noe et al. PNAS (2009)

Trajectory Mapping

Top-down

Clustering MD trajectories to form metastable states

- Decrease size of data set in clustering
- Depress intrastate fluctuations very much but keep the interstate fluctuations
- Take into account the similarity on dynamics

Trajectory Mapping

- Mapping each MD trajectory to a high-dimensional vector to represent the configuration distribution in the trajectory

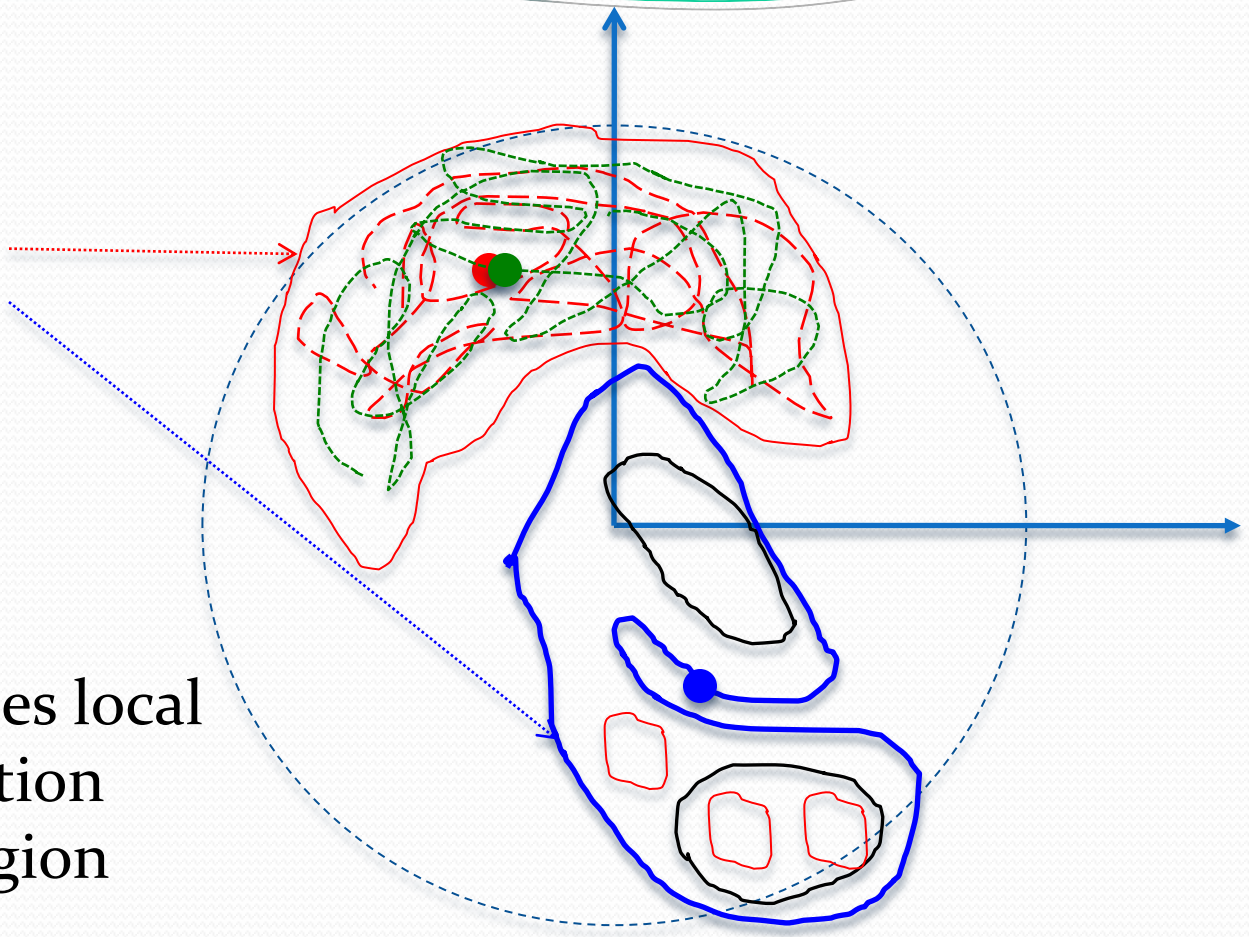
$$q(t \in [0,\tau]) \rightarrow \vec{v} = (< A^1(q) >, ..., < A^n(q) >)$$

$$< A^\mu(q) >= \frac{1}{\tau}\int_0^\tau dt A^\mu(q(t)) = \int dq A^\mu(q) P(q)$$

$\{A^i(q)\}$ is a set of functions of configuration (basis functions)

Trajectory Mapping

Shape and size of metastable region may be complicate



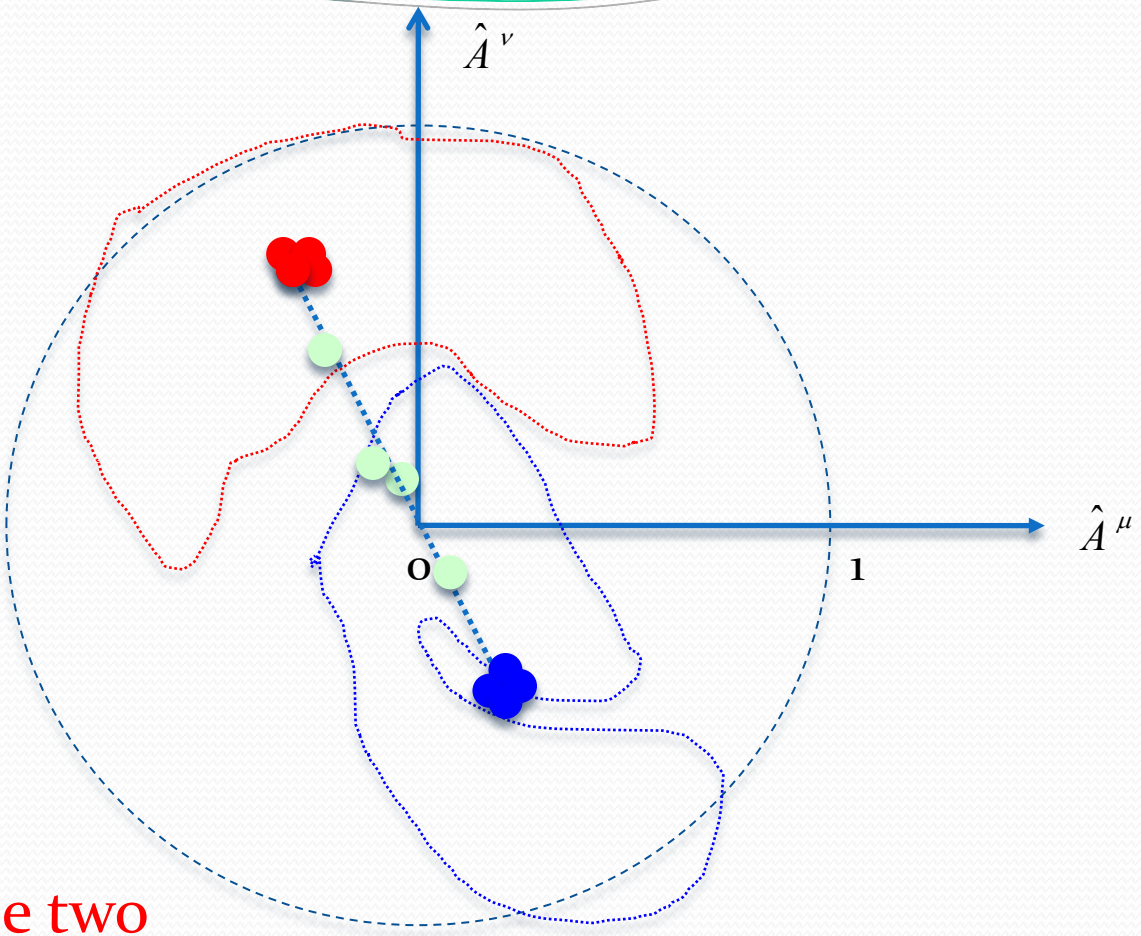
MD Trajectory reaches local equilibrium distribution inside metastable region

the trajectory-mapped vectors are almost same
(the center limit theorem)

A tau-length trajectory

locates inside a tau-scale meta-stable state and reach local equilibrium

or transition among a few states and reach local equilibrium in each of these meta-stable states

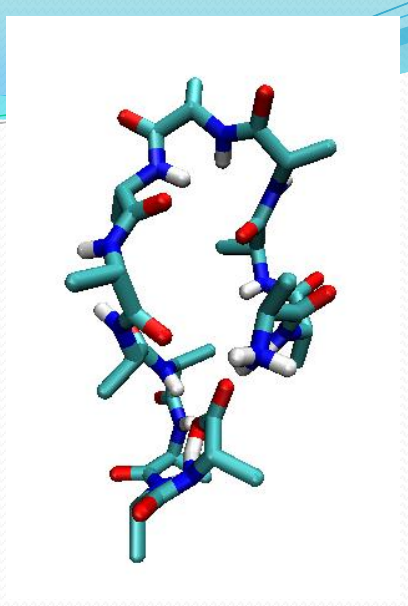


All trajectories inside two states are mapped in a line

Trajectory-mapped vectors have much simpler geometry

Trajectory Mapping

- 74 atoms, charged terminals;
- Implicit solvent simulation: Generalized Born;
- 172 basis functions from torsion angles

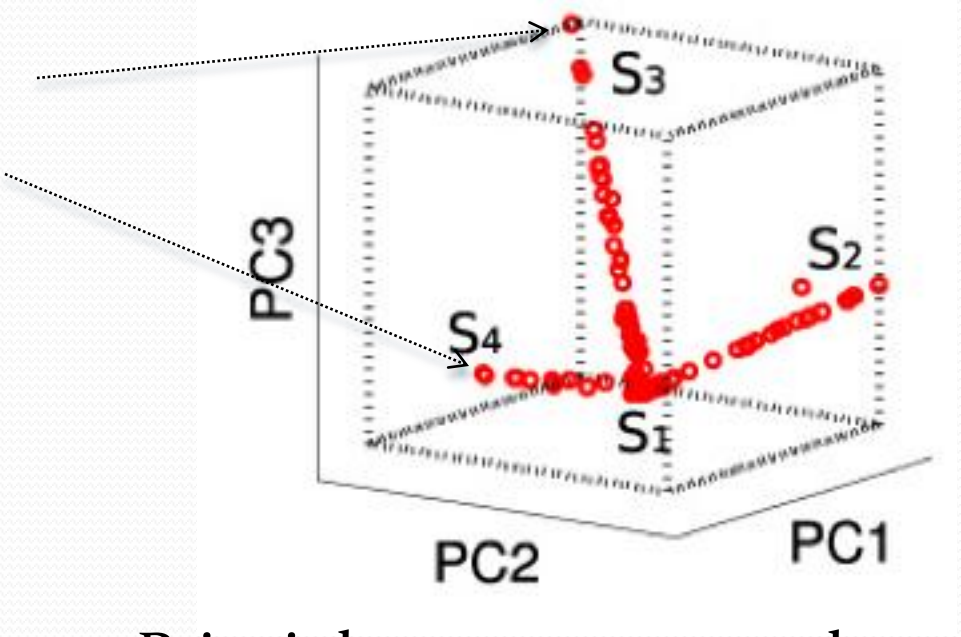


Alanine-dodeca-peptide

Each 1ns-length MD trajectory is mapped to a 172d vector

Clustering

Four nanosecond-order metastable states are found



Principle component analyse

Orthonormalized basis functions

$$\frac{P(x)}{P_{ref}(x)} = 1 + \sum_{\mu} \left\langle \hat{A}^{\mu}(x) \right\rangle_{P(x)} \hat{A}^{\mu}(x) + \ldots$$

$$\left\langle \hat{A}^{\mu}(x) \right\rangle_{P_{ref}(x)} = 0$$

$$\left\langle \hat{A}^{\mu}(x) \hat{A}^{\nu}(x) \right\rangle_{P_{ref}(x)} = \delta_{\mu\nu}$$

A trajectory is mapped to a n-dimensional vector

$$q(t) \rightarrow P(q) \rightarrow \vec{v} = (< \hat{A}^{1}(q) >, \ldots, < \hat{A}^{n}(q) >)$$

Inner product between trajectories is related to their overlapping

$$\vec{P}_i \circ \vec{P}_j \equiv \int dx \, \frac{P_i(x) P_j(x)}{P_{ref}(x)} \approx 1 + \sum_{\mu} \left\langle \hat{A}^{\mu}(x) \right\rangle_{P_i(x)} \left\langle \hat{A}^{\mu}(x) \right\rangle_{P_j(x)} = 1 + \vec{v}_i \bullet \vec{v}_j$$

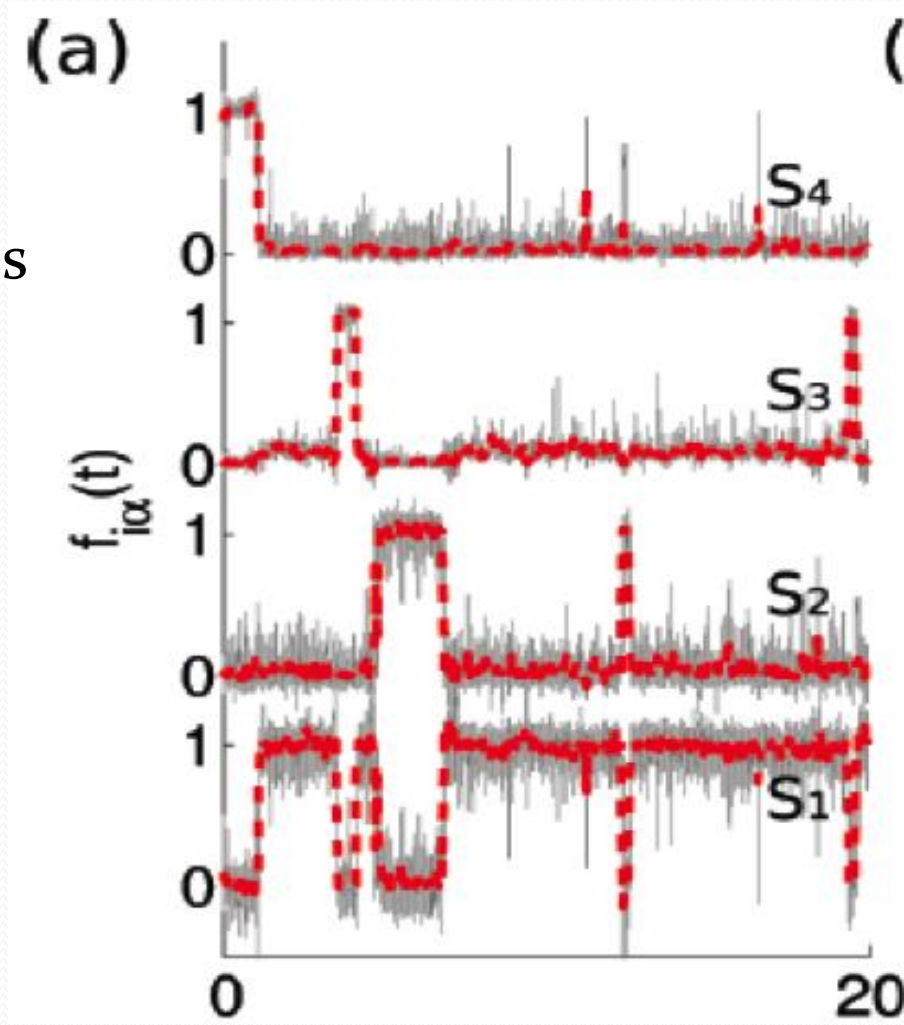
$$f_\alpha(t) \equiv |P^\alpha(x)|^{-2} \int dx \frac{P^\alpha(x)\delta(x - x(t))}{P_{ref}(x)} = \begin{cases} 1, & x(t) \in S_\alpha \\ 0, & otherwise \end{cases}$$

$$P_{ref}(x) = \sum_\alpha c_\alpha P^\alpha(x)$$

Calculating inner product of single configuration with meta-stable states

f(t) is state-indicator curve of trajectory

Transition kinetics among states



(c)

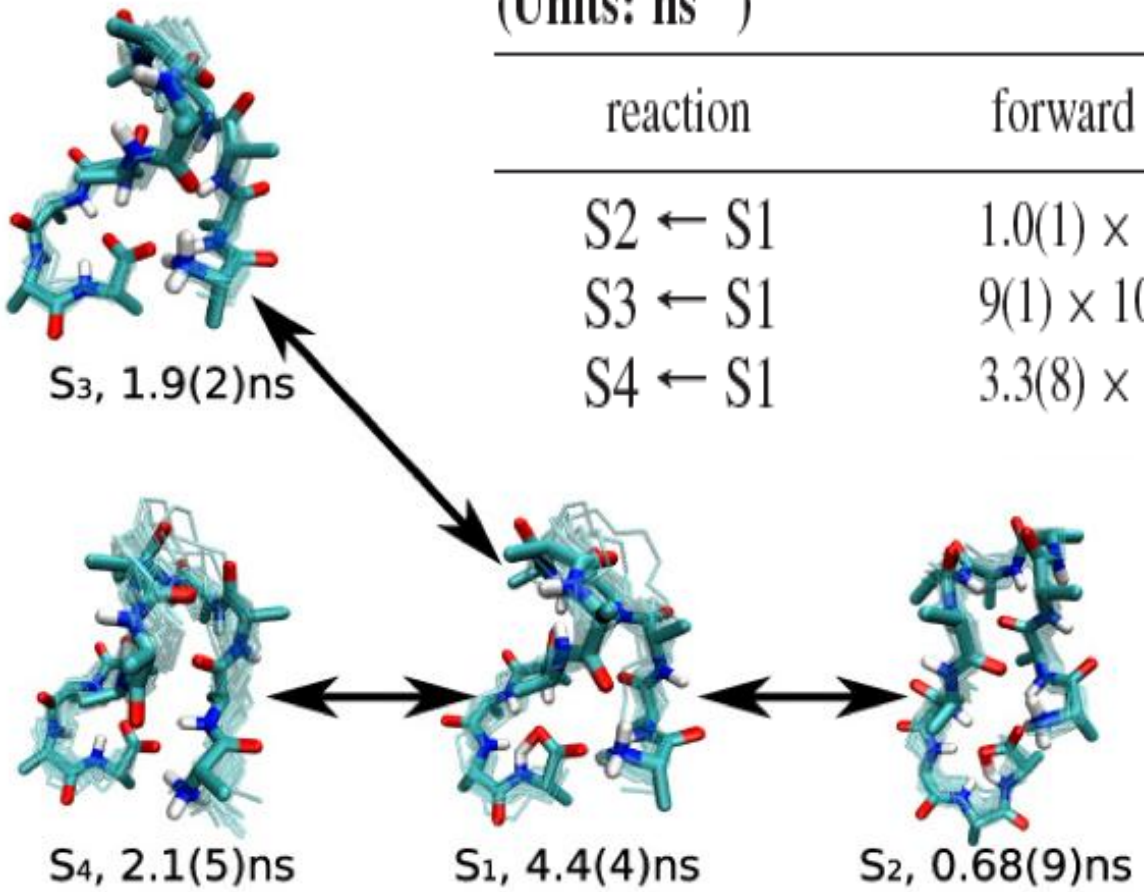
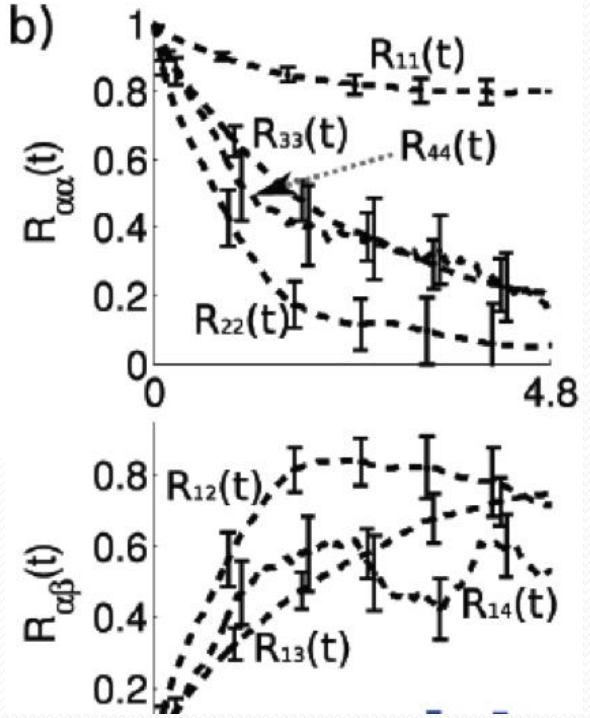


TABLE 1: Kinetic Rates of Four-State Transition Network (Units: ns^{-1})[a]

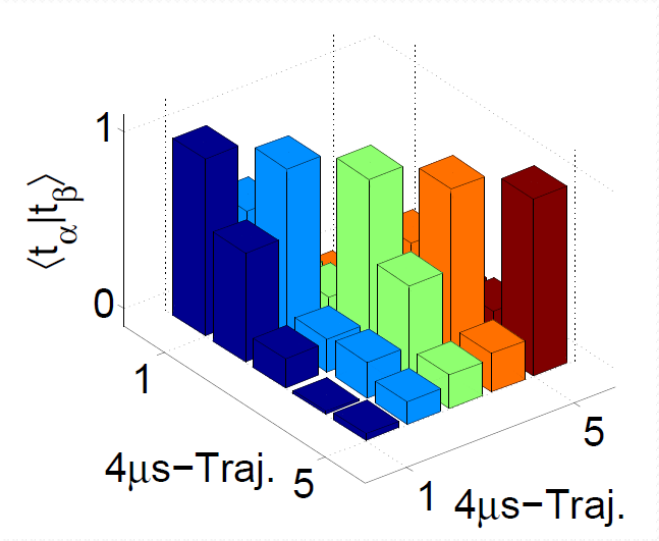
reaction	forward rate	reverse rate
S2 ← S1	$1.0(1) \times 10^{-1}$	1.4(2)
S3 ← S1	$9(1) \times 10^{-2}$	$5.2(7) \times 10^{-1}$
S4 ← S1	$3.3(8) \times 10^{-2}$	$4(1) \times 10^{-1}$

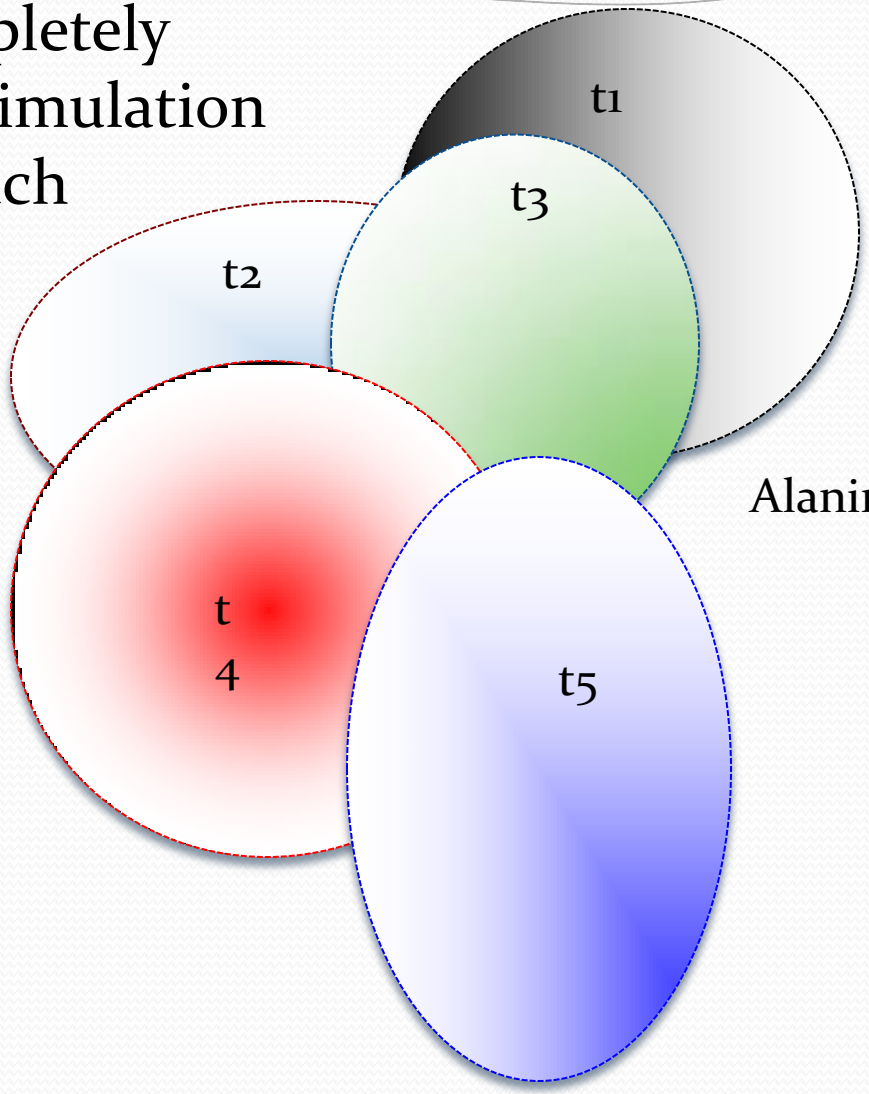


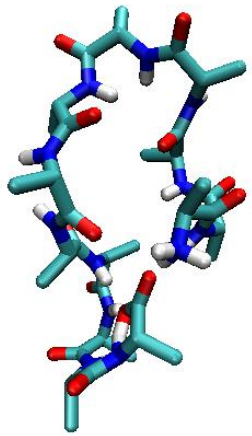
Transition kinetics

L. Gong and XZ, JPCB (2010)

Five 4-us trajectories overlap partially but not completely (microsecond-order simulation is not sufficient to reach equilibrium)



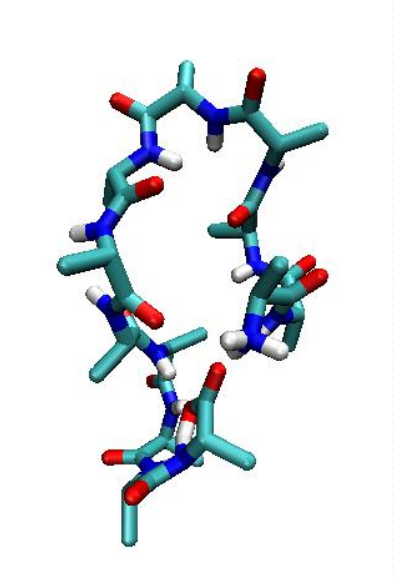




Alanine-dodeca-peptide

$$\vec{P}_i \circ \vec{P}_j \equiv \int dx \, \frac{P_i(x) P_j(x)}{P_{ref}(x)} = 1 + \sum_{\mu} \left\langle \hat{A}^{\mu}(x) \right\rangle_{P_i(x)} \left\langle \hat{A}^{\mu}(x) \right\rangle_{P_j(x)} = 1 + \vec{v}_i \bullet \vec{v}_j$$

Folding network of Ala_{12}

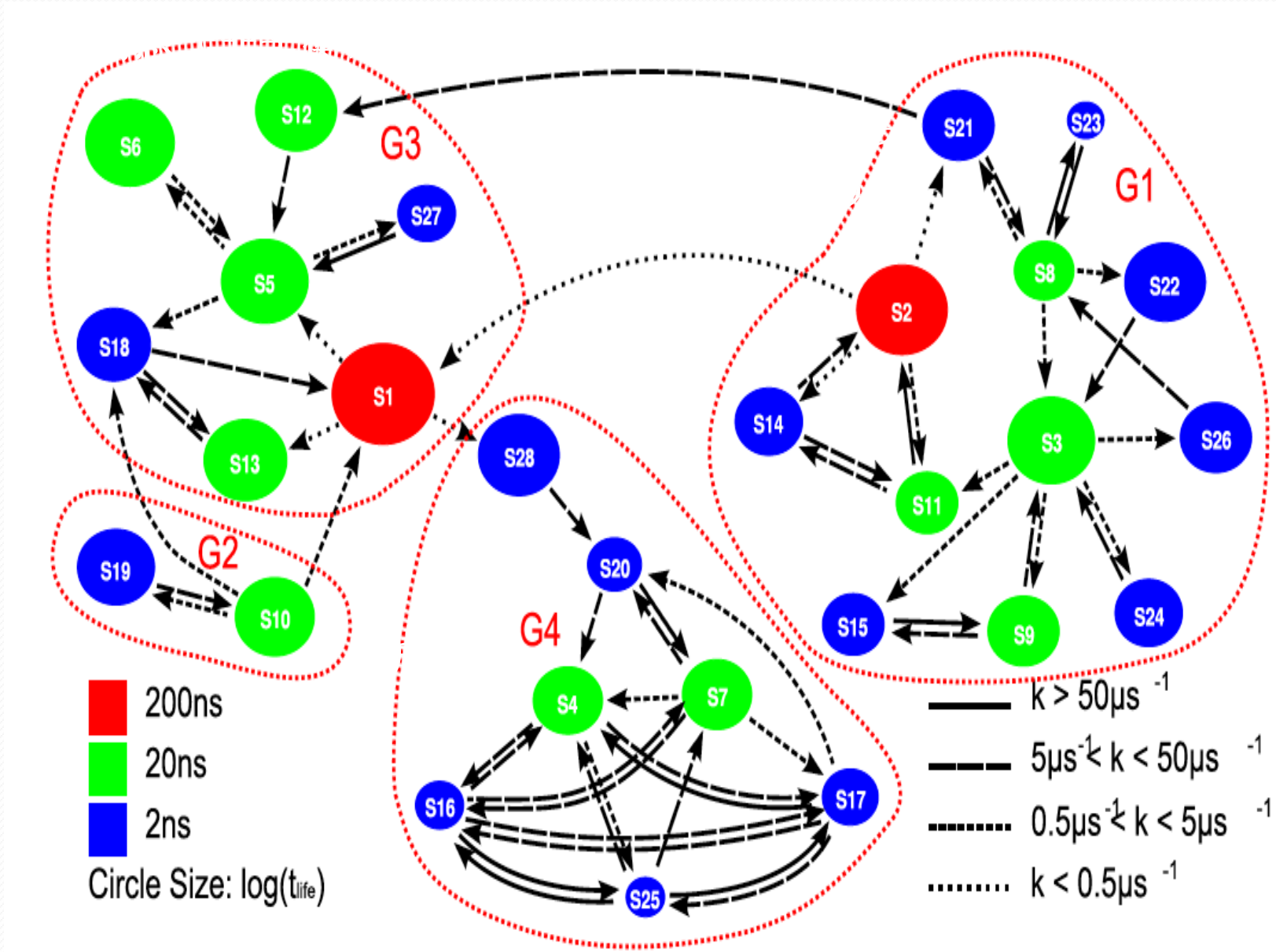


Analysis

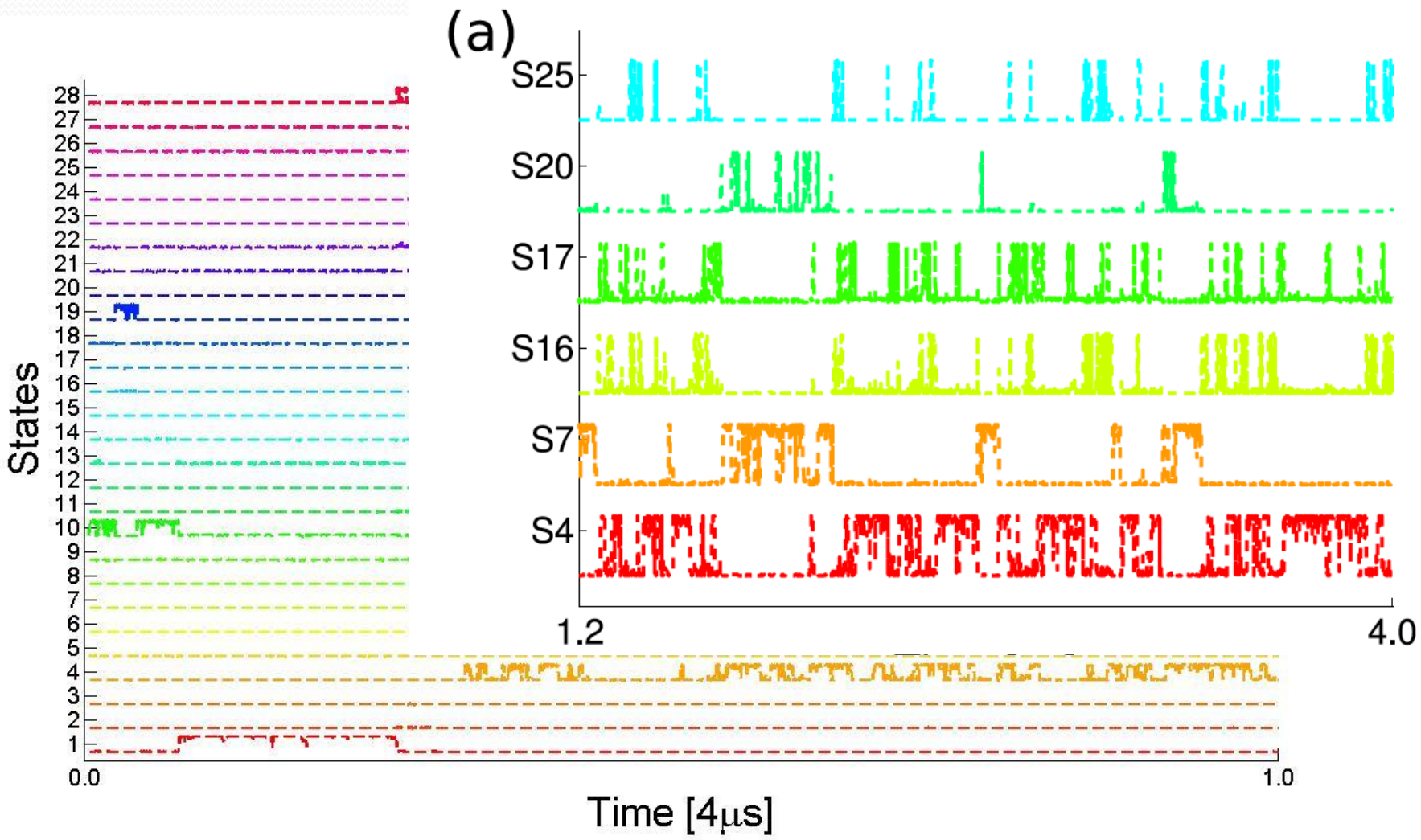
- Three levels: 200ns, 20ns, 2ns;
- 28 states found, accounting for more than 90% simulation data.

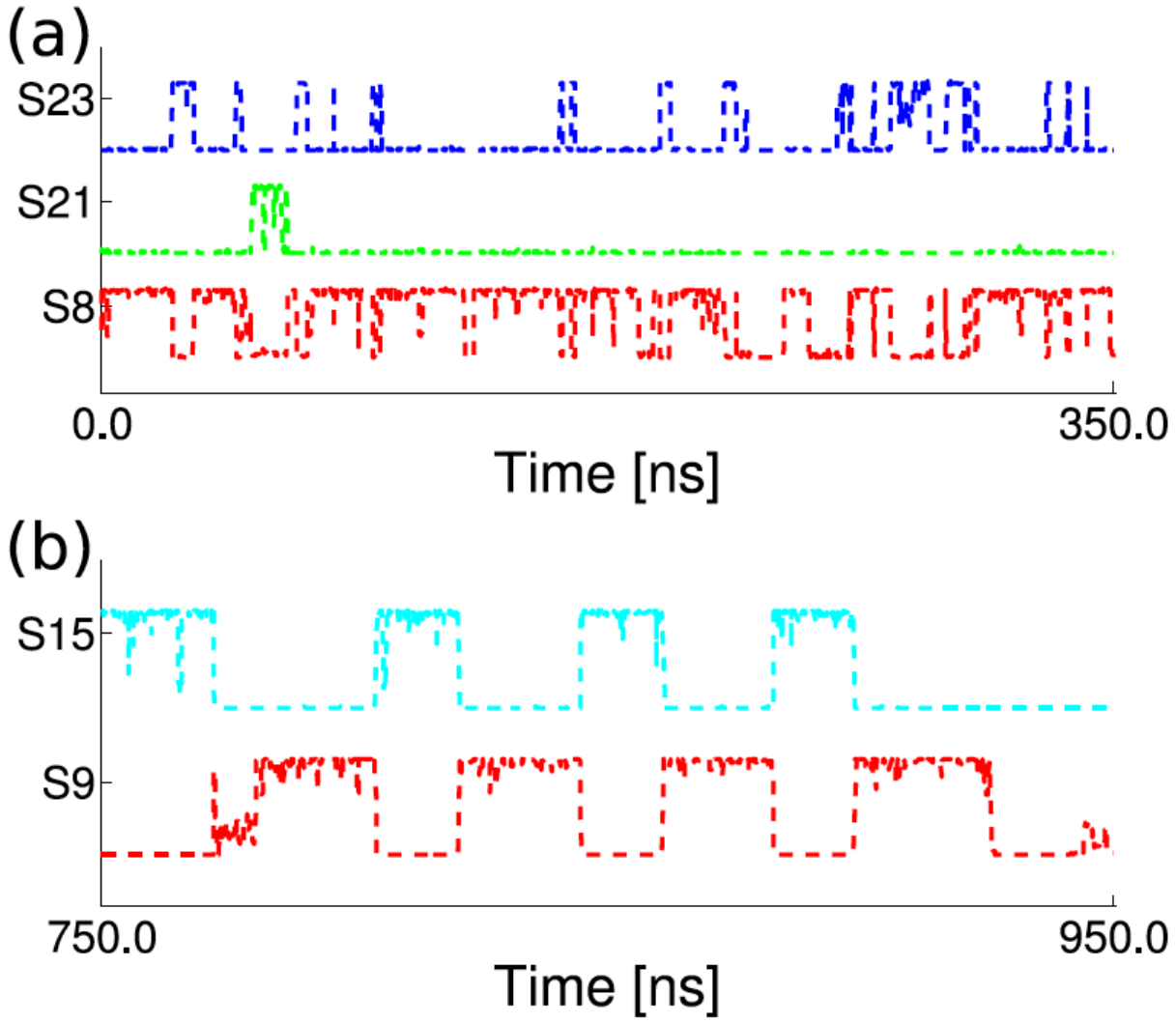
Get meta-stable states from long to short time scales

Metastable state network of Ala_{12}

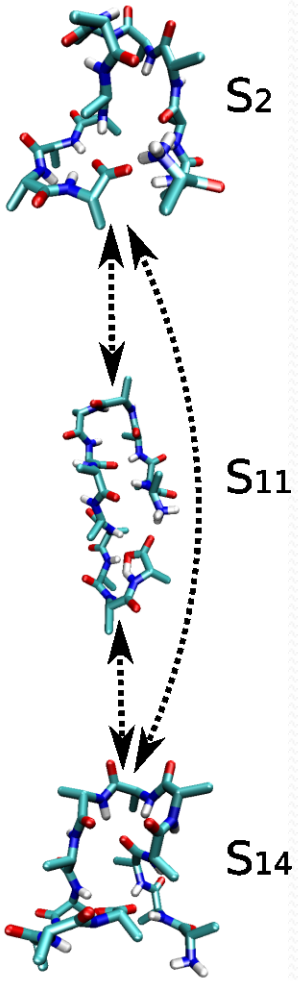


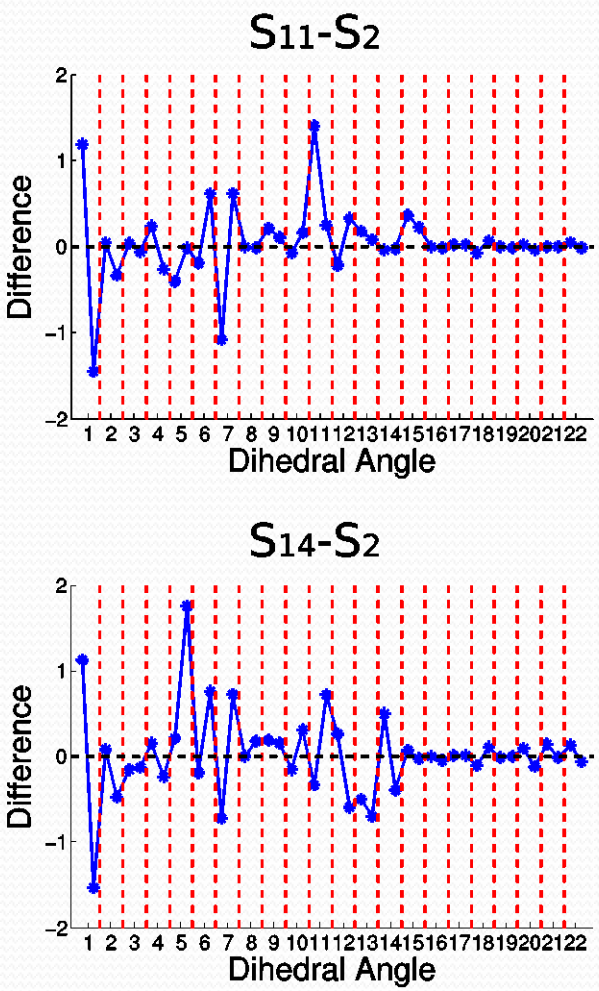
Transition trajectory

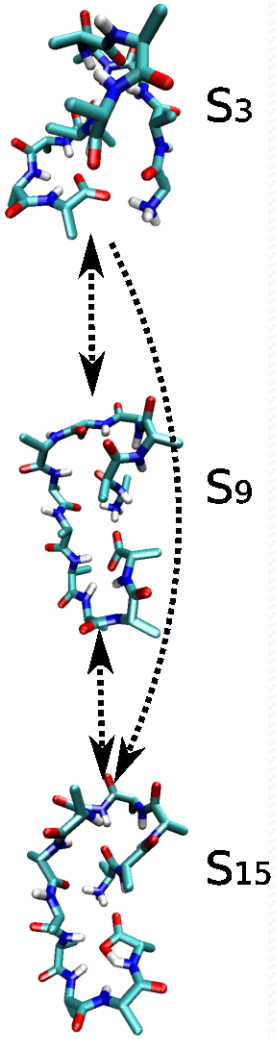


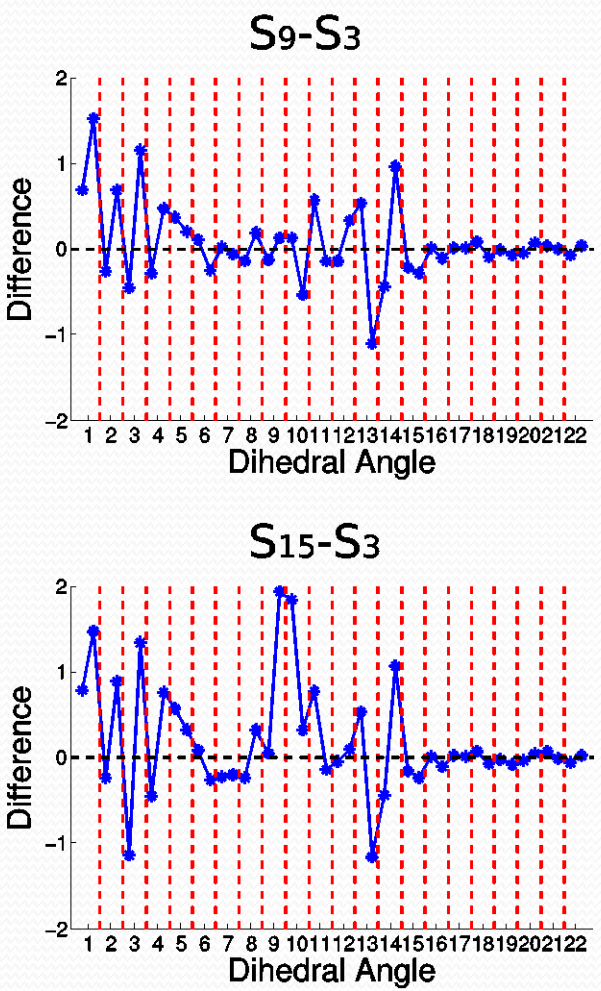

(a)
S23
S21
S8
0.0
350.0
Time [ns]
(b)
S15
S9
750.0
950.0
Time [ns]

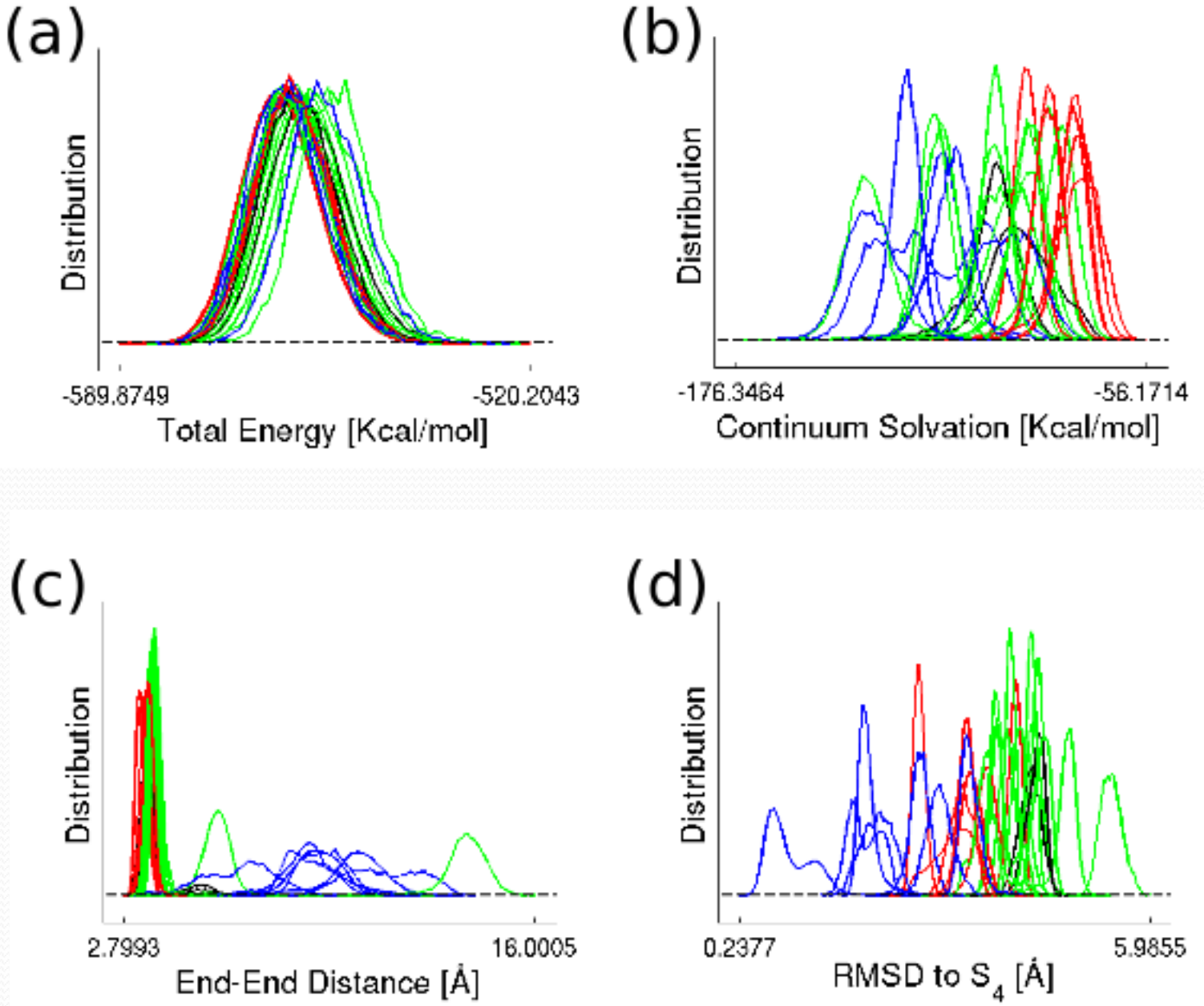
Reaction Detail Derived from Folding Network (Main Chain Dihedral Angle Difference)










(a)
Distribution
-589.8749
-520.2043
Total Energy [Kcal/mol]
(b)
Distribution
-176.3464
-56.1714
Continuum Solvation [Kcal/mol]
(c)
Distribution
2.7993
16.0005
End-End Distance [Å]
(d)
Distribution
0.2377
5.9855
RMSD to S4 [Å]

Summary

Trajectory mapping and clustering identify metastable states in high-dimensional configuration space

Metastable states are dependent on dynamics and time scale

$$t_{eq} / t_{life} << 1$$

Transition kinetics among metastable states may be achieved or be focused on (e.g. transition path sampling, string method, flux method, etc.)

x: Collective Variable

$$\{q_1,\ldots,q_M\} \rightarrow p(x) \rightarrow \vec{v} = (\langle \hat{A}^1(x)\rangle,\ldots,\langle \hat{A}^n(x)\rangle)$$

Difference between samples

$$d(P_i,P_j) \equiv \int dx \frac{|P_i(x)-P_j(x)|^2}{P_{ref}(x)}$$

$$\approx |\vec{v}_i - \vec{v}_j|^2$$

$$\frac{P(x)}{P_{ref}(x)} = 1 + \sum_\mu \langle \hat{A}^\mu(x)\rangle_{P(x)} \hat{A}^\mu(x) + \ldots$$

$$\langle \hat{A}^\mu(x)\rangle_{P_{ref}(x)} = 0$$

$$\langle \hat{A}^\mu(x)\hat{A}^\nu(x)\rangle_{P_{ref}(x)} = \delta_{\mu\nu}$$

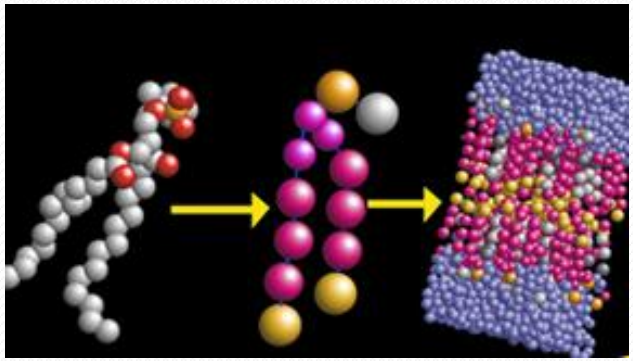
$$\langle \hat{A}^\mu(x)\rangle \approx \frac{1}{M}\sum \hat{A}^\mu(x_i)$$

Coarse-graining

Coarse-graining

1. Map CG dof : x=X(r)
2. Select effective potential formula

$$U(x;u_\lambda) = \sum u_\lambda f^\lambda(x)$$



3. Optimize parameters of U(x) by minimizing difference between U(x) and F(x)

$$F(x) \equiv -k_B T \ln \int dr e^{-\beta V(r)} \delta(x - x(r))$$

Coarse-graining: Match probability density

$$d^2(F,U) = \left\langle \left(\frac{P_{cg}(x) - P_{aa}(x)}{P_{ref}(x)} \right)^2 \right\rangle_{ref}$$

$$\approx \sum (g^{-1})^{\mu\nu} a^{\mu} a^{\nu}$$

$$a^{\mu} = < A^{\mu} >_{cg} - < A^{\mu} >_{aa}$$

$$\left\langle A^{\mu}(x) A^{\nu}(x) \right\rangle_{P_{ref}(x)} = g^{\mu\nu}$$

Covariance matrix

Correlation among of thermodynamics variables should be removed

$$P_{ref}(x) = \frac{P_{cg}(x) + P_{aa}(x)}{2}$$

$$d^2(U,F) = \sum_{\mu} (\hat{a}^{\mu})^2$$

Relationship of different CG Matching

$$d^2(U,F) = \sum_{\mu} (\hat{a}^{\mu})^2$$

Corrected thermodynamics matching

$$e^{U(x)} \sim e^{F(x)}$$

Free energy surface (or probability) matching

$$\nabla U(x) \sim \nabla F(x)$$

Gradual of free energy surface matching

$$\nabla U(x) \sim \langle \nabla V(r) \rangle_x$$

$$x = \sum_{\alpha} a_{\alpha} r_{\alpha}$$

Force matching (with linear CG transformation)

$$d^2(U,F) = \sum_{\mu=1,\ldots m} (\hat{a}^{\mu})^2$$

- In principle, m must be infinite

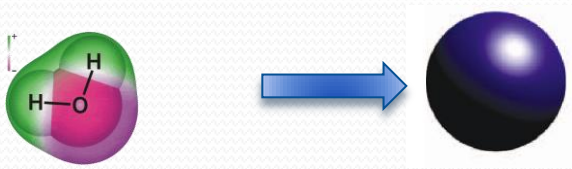
- In practice, m << M, the size of sample
(the improvement using more basis functions is not helpful due to larger statistical error)

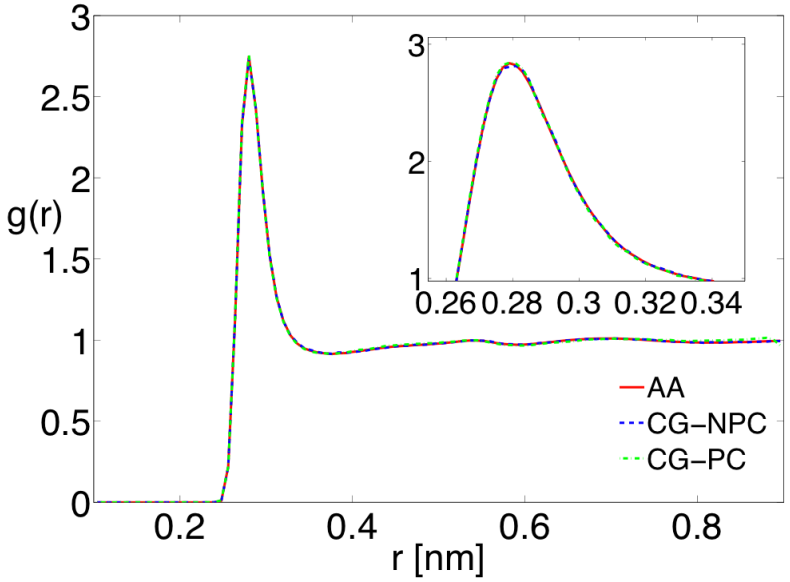
- The upper limit of relative deviation of any <A(x)> in the coarse-graining is d(U,F)

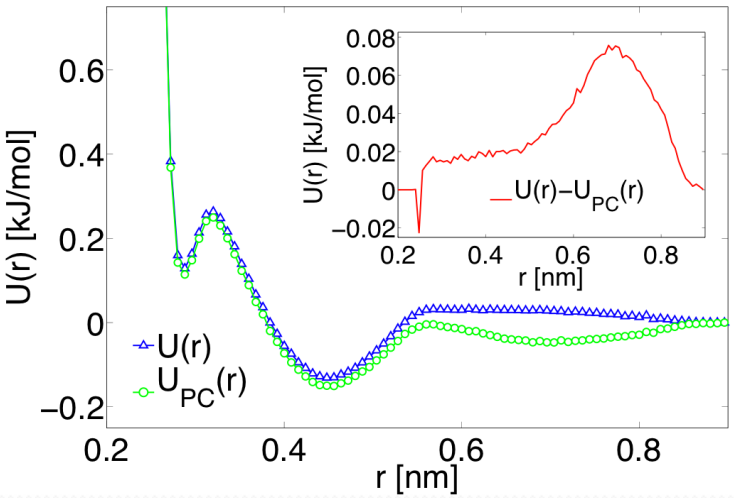
$$\left| \langle A(x) \rangle_{cg} - \langle A(x) \rangle_{aa} \right| \quad < \quad \sigma_{ref}(A)\, d(U,F)$$

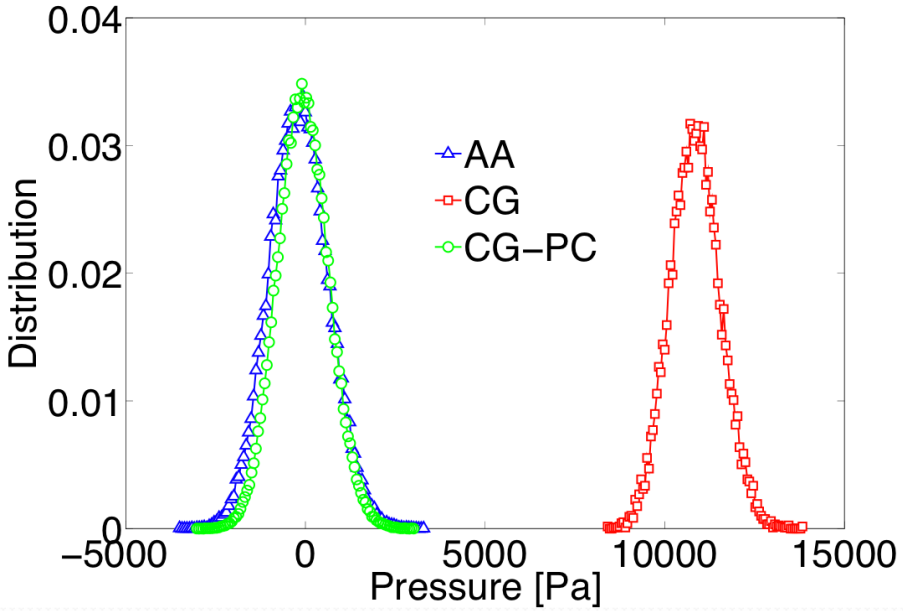
$$d^2(U,F) = \sum_{\mu} (\hat{a}^{\mu})^2$$

one-site CG water model







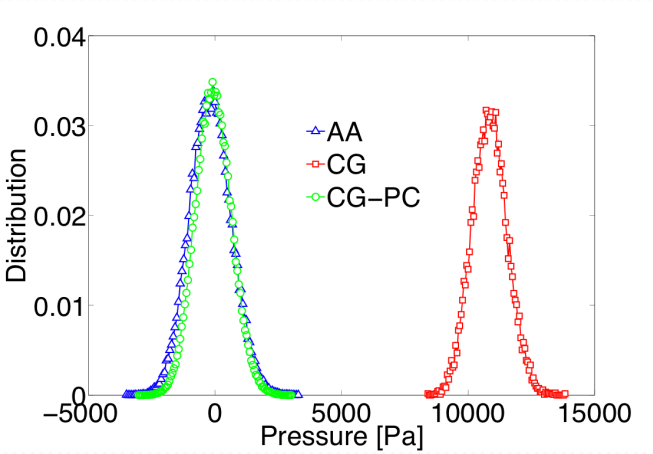


Pressures in CG and AA are not the ensemble means of the same configuration function

$$P_{cg} = \left\langle \hat{P}_{cg}(x) \right\rangle_{cg}$$

$$P_{aa} = \left\langle \hat{P}_{aa}(r) \right\rangle_{aa}$$

$$\hat{P}_{cg}(x(r)) \neq \hat{P}_{aa}(r)$$



Matching probability density in extended space:

$$d(P_{cg}(\vec{x}^n), P_{aa}(\vec{x}^n))$$

$$P_{cg}(\vec{s}^n; V) \sim P_{aa}(\vec{s}^n; V)$$

$$\vec{s} = \vec{x} / L$$

Summary

The mean of any CG-configuration function A(x) can be reconstructed in probability density matching CG

Matching probability density in an individual canonical ensemble is not sufficient for reconstruction of pressure (and some another physical variables, such as E, chemical potential)

The matching method is actually a reverse MC. Its transferability was not be guaranteed.

Question: how to construct a more transferable coarse-graining model?

Kavli Institute of Theoretical Physics, China (KITPC)

Special Program on

Advanced Molecular Simulation Methods in the Physical Sciences

Bejing, China, June 10 – July 5, 2013

http://kitpc.itp.ac.cn/program.jsp?id=PA20130610

- **Trajectory space sampling**
- **Weighted ensemble sampling**
- **Coarse-graining methods**
- **Free energy calculations**
- **Molecular simulation of self-assembly**
- **… …**

Welcome

Thank you for your Attention!

xzhou@gucas.ac.cn

School of Physical Sciences
Graduate University of Chinese Academy of Sciences

中國科學院研究生院
GRADUATE UNIVERSITY OF CHINESE ACADEMY OF SCIENCES