University of Tsukuba 筑波大学

CCS-LBNL collaborative joint workshop
Tsukuba Japan, May 12th -13th 2016

Efficient Sampling Methods for Rare Events by Molecular Dynamics Simulations

Ryuhei Harada and Yasuteru Shigeta

Division of Life Sciences

University of Tsukuba, Center for Computational Sciences (CCS)

Acknowledgement

Collaborators

Hiroshima City University/ Osaka University

Dr. Y. Takano (Prof: New sampling method)

Fujitsu Co. Ltd.

Dr. T. Nakamura (Clustering Algorithms)

Osaka Univ.

Dr. T. Baba (previous student: Molecular dynamics simulation and enzymatic reaction of Nylonoligomer hydrolase)

H. Ando (previous student: Interaction analyses of Nylonoligomer hydrolase)

Dr. M. Nakano (Prof.: Discussion on Nylon-oligomer hydrolase)

University of Hyogo

Dr. S. Negoro (Prof: Biochemical analyses on Nylonoligomer hydrolase)

Dr. Y. Higuchi (Prof: Xray structural analyses on nylonoligomer hydrolase)

Dr. N. Shibata (Prof: Xray structural analyses on nylonoligomer hydrolase)

Strasbourg Univ./CNRS/IPCMS

Dr. M. Boero (Prof :Metadynamics Simulation using CPMD

Funding

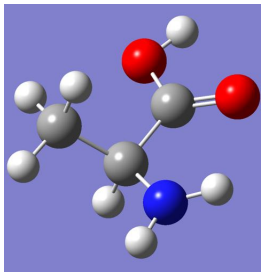
○Grant-in-Aid for the scientific research on Innovative area MEXT, Japan
"3D active site" , "Dynamic ordering and functions", and
"Photosymergetics" (Nos. 26107004 , 26102525, and 26105012).

○CREST Program "Post peta-scale computer architecture" from JST.

Complexity and Hierarchy of Life Sciences

A system in a different scale requires a different physics law

Amino Acids
1nm



L-Alanine

20+α kinds

Our target

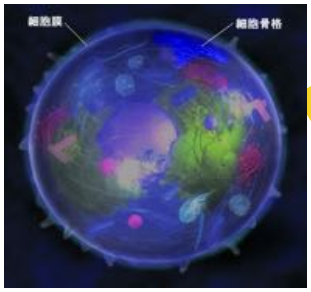
Proteins
10-100nm

Cytochrome c

~10^5 kinds

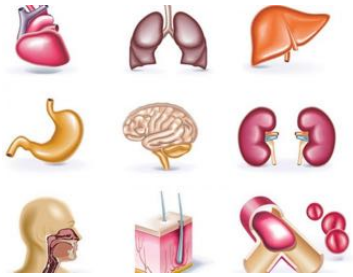
10^4~10^6 atoms

Cells
1–100μm



~300 kinds

Organs
1–10cm



~10 kinds

Human
1m

~$6x10^{12}$ cells

Wealth of Biological Functions

- ○ **Transport processes**
- ○ **Energy conversions**
- ○ **Molecular Recognitions**
- ○ **Signal transductions**
- ○ **Enzymatic reactions**

Toward Analyzing Protein Functions

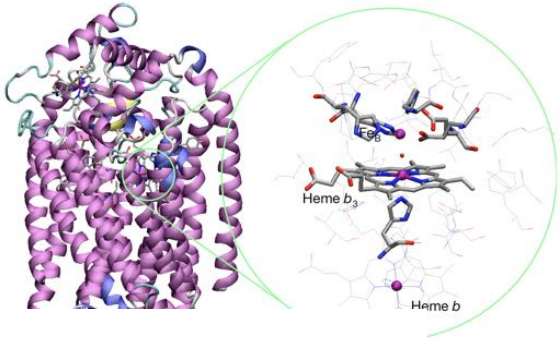
Accuracy

Quantum Mechanics

Fragment MO

Analyses on Interaction energy

RSCPMD/PWCPMD



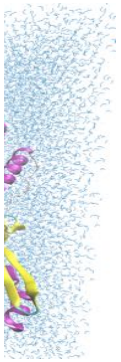
Our group performs

Hybrid Methods

Methodological development

&

Biological applications

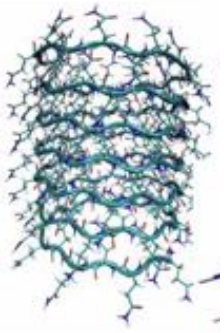


Analyses on structures, large amplitude motion, etc.

Classical Mechanics

Coarse-grained/ Accelerated MD

Protein-Protein, Protein-Enzyme complex formation



Bio informatics

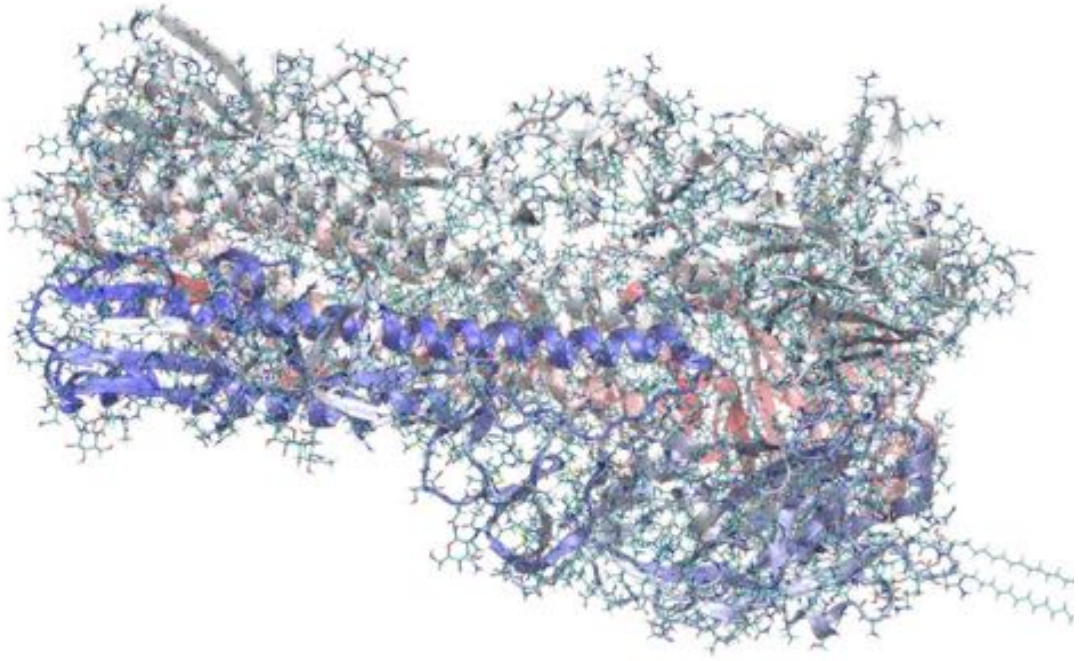
GPU accelerated Fragment Molecular Orbital

Key Paper: H. Umeda, H. Hanawa, M Shoji, T. Boku, YS, *Journal of Computer Chemistry Japan* (**invited letter**) **14**, 69-70(2015).

- **Influenza HA3 Protein (23,460 atoms, 721 fragments)**
- FMO-HF/6-31G(d)
- HA-PACS base cluster 64nodes
  - 1024 CPU core + 256 GPU
- 84 worker group
  - 3 MPI rank/worker group

- **11 min** by K Computer 24,576 nodes (**3.1PFLOPS**)
- **120 min** by HA-PACS base cluster 64 nodes(**192TFLOPS**)

	HA3
#nodes (#GPU)	64 (256)
SCC [hr]	0.52
Dimer SCF [hr]	0.90
ES Dimer [hr]	0.45
Total [hr]	1.97



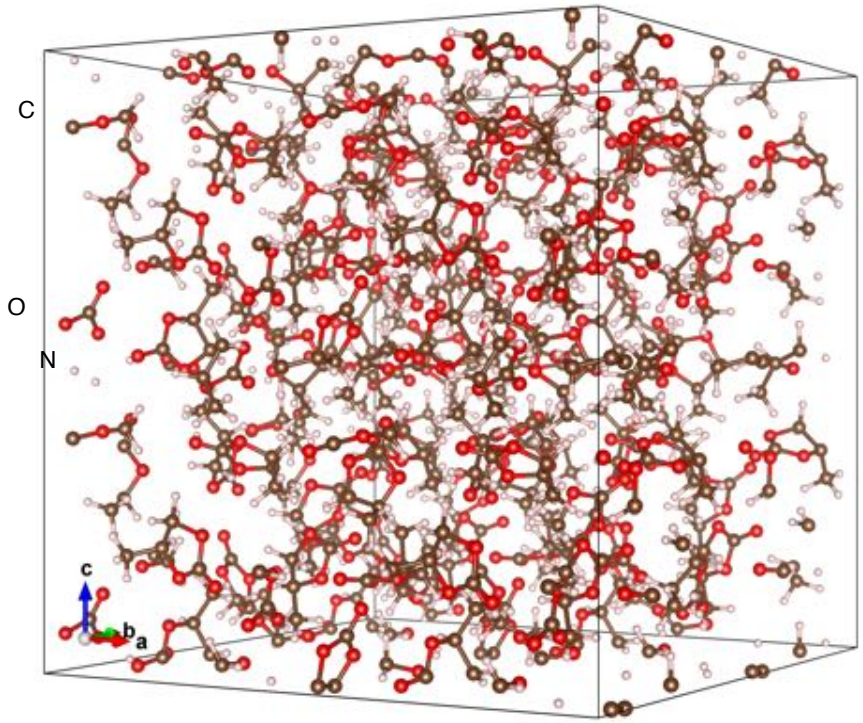
GPU accelerated Car-Parrinello MD

Molecular fluid (1664 atoms)
# of grid points : 160×160×160 = 4,096,000
（～ 103 Ry ）
# of bands : 2560

MPI_WTIME (sec.)

PW-CPMD	1 MD step
512 MPI × 8 OMP (4096 cores) on K	11.9

CPMD (http://www.cpmd.org) tuned for K computer



MPI_WTIME (sec.)
(FLOPS/PEAK %)

Selected as one of post K computer projects

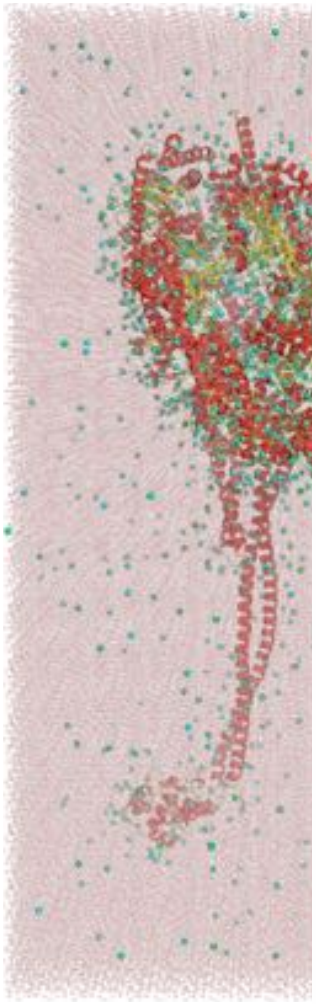
RS-CPMD	1 MD step	rotorb	Hamiltonian & Force	Hamiltonian operation	rotorb2
512 MPI × 8 OMP (4096 cores) on K	5.54 (37.5 %)	2.69	0.52	0.44	1.83

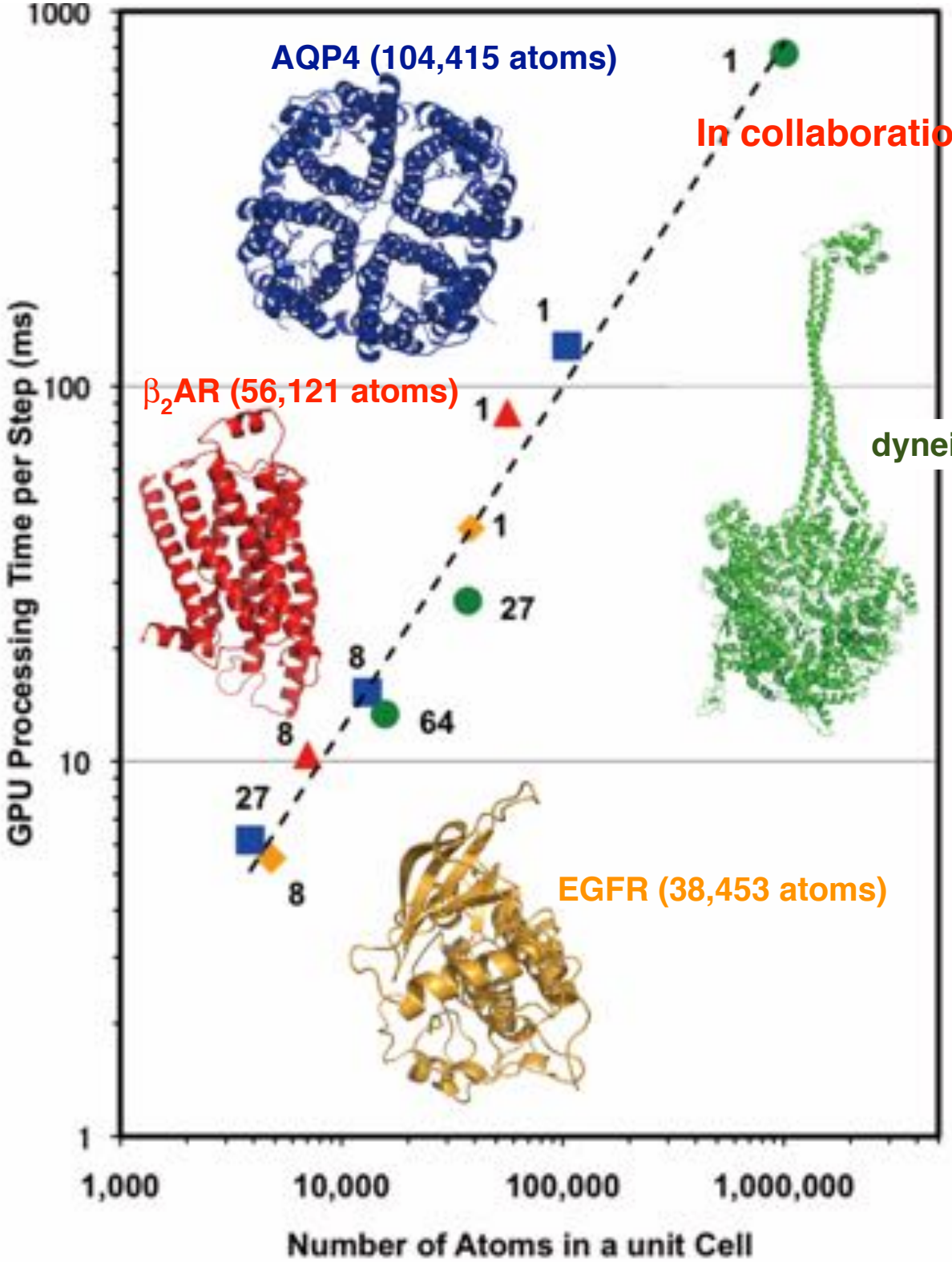
By using super computer at Kashiwa (512 MPI × 1 OMP 128 GPUs/64 nodes) both rotorb and rotorb2 become 3.4 times faster than CPU version on average

GPU … ics (MD)

dynein(
574Na+

689 TIP3P water +
…atoms





In collaboration with Osaka Univ.

…o even 125GPUs

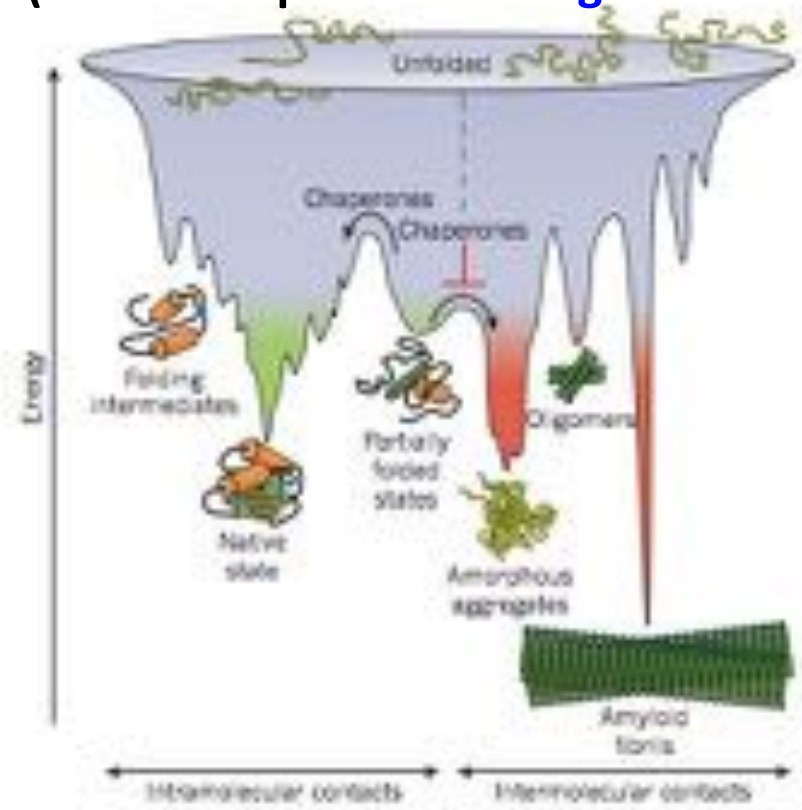
…times faster

comm
44.8
15.6
6.7
4.5

Protein Folding as Rare Events

◆**Protein folding**

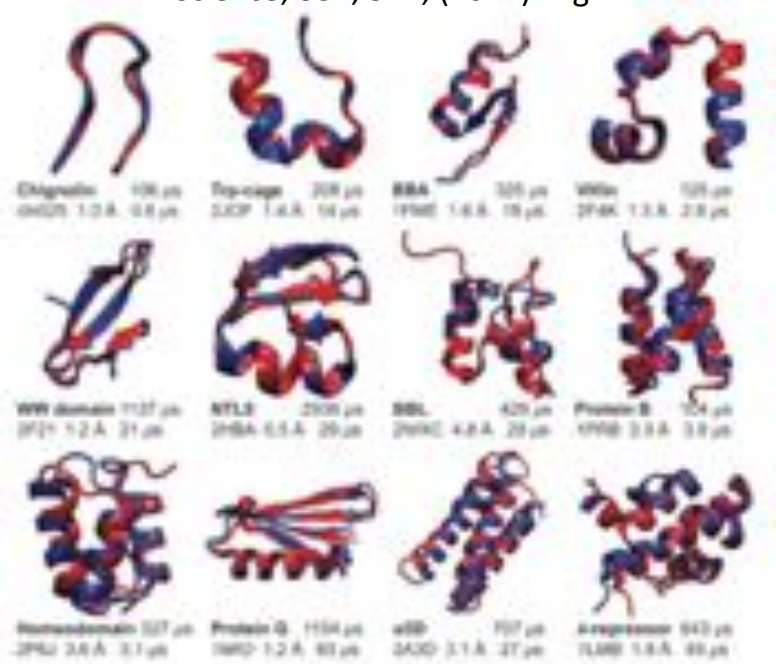
(**Stochastic process in long-time scale**)



Nature, **475**, 326 (2011). Fig. 1

◆**ANTON @ D. E. Shaw Research**

Science, **334**, 517, (2011). Fig. 1

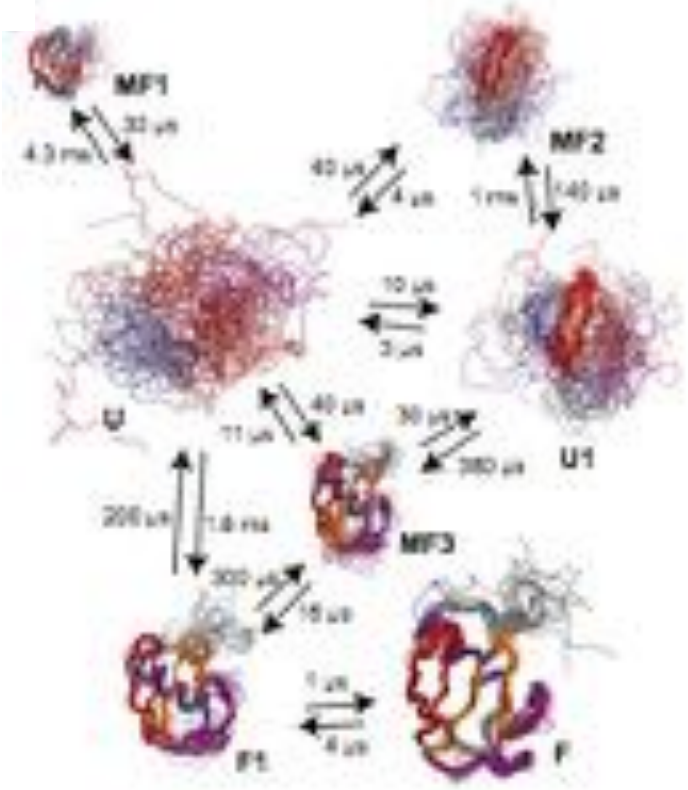


To extract biologically important rare events, more than microsecond to second order molecular dynamics (MD) simulations are required!

Brute Force vs. Distributed Computing

◆**Brute force** MD simulations by ANTON

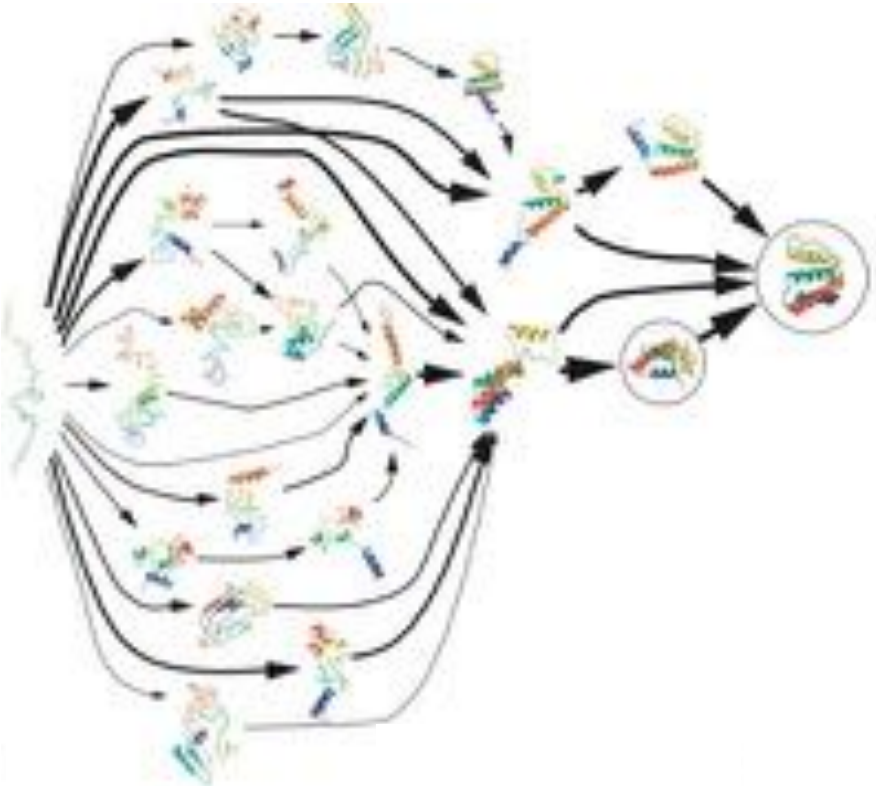
Millisecond order all-atom folding simulation of ubiquitin by ANTON. (*PNAS*, **110**, 5915, (2013). Fig. 1)



versus

◆**Distributed computing** by multiple short-time MD simulations

Folding of ACBP represented by Markov state model. (*Curr. Opin. Struc. Biol.* **23**, 58, (2013). Fig. 3)



Only one long-time trajectory

Time-series: ○

Statistical ensemble: ×

Parallel Processing : △

Multiple short-time trajectories

Time-series: ×

Statistical ensemble: ○

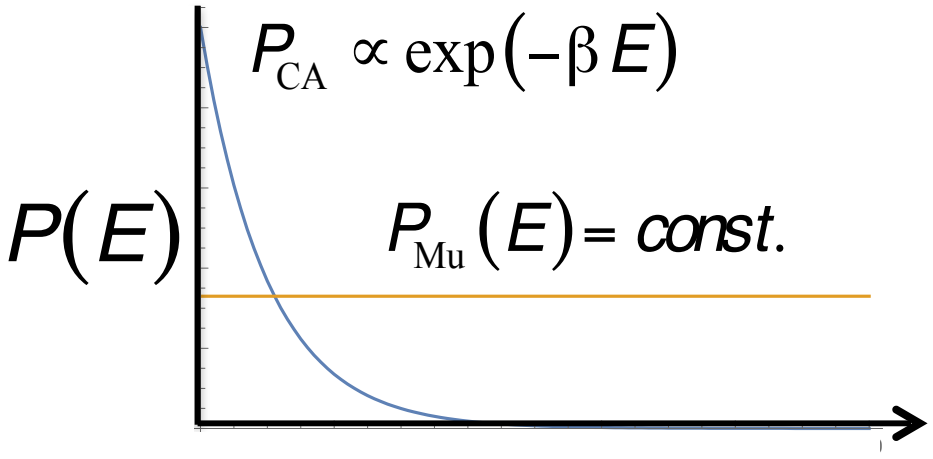
Parallel Processing : ○

Smart Ways: Generalized Ensemble

1. Multi-canonical Ensemble Method

Details are discussed in Lecture by Prof. Okamoto

B.A. Berg, T. Neuhaus, *Phys. Lett.* **B267**, 249 (1991).
B.A. Berg, T. Neuhaus, *Phys. Rev. Lett.* **68**, 9 (1992).
U.H.E. Hansmann, Y. Okamoto, F. Eisenmenger, *Chem. Phys. Lett.*, **259**, 321 (1996).



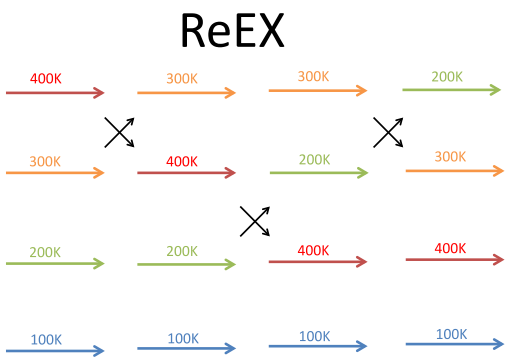
Non-Boltzmann weight

$$P_{Mu}(E) \propto n(E) W_{Mu}(E) = const.$$

Reweighting method to reproduce canonical ensemble

2. Replica Exchange Method

Y. Sugita, Y. Okamoto, *Chem. Phys.Lett.* **314**, 141 (1994)
K. Hukushima, K. Nemoto, *J. Phys. Soc. Jpn.* **65**, 1604 (1996)
K. Hukushima, H. Takayama, K. Nemoto, *Int. J. Mod. Phys. C* **7**, 337 (1996).

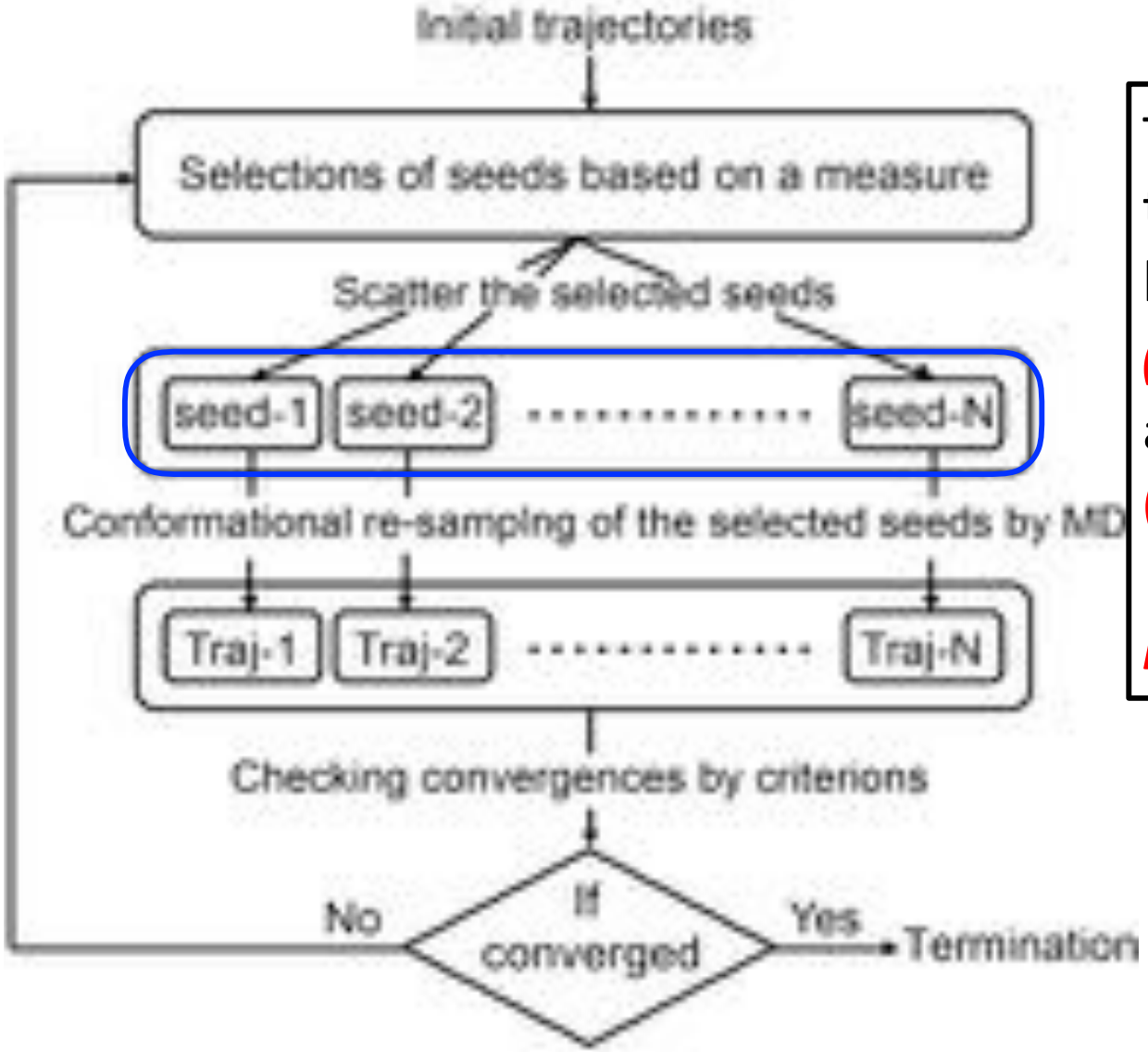


3. Simulated Tempering Method

A.P. Lyubartsev, *et al.*, *J. Chem. Phys.* **96**, 1776 (1992).
E. Marinari and G. Parisi, *Europhys. Lett.* **19**, 451 (1992).

Basic Concept of Our Methods Based on Distributed Computing

R. Harada, Y. Takano, T. Baba, **YS**, *Phys. Chem. Chem. Phys.* (***Feature Article***) **17**, 6155-6173 (2015)



The basic concept for enhancing the conformational sampling is based on

(i) selections of seeds

and

(ii) ***conformational re-sampling from them, after re-generation of velocities***.

Multiple short-time MD simulations are performed on a large amount of individual GPUs for conformational re-sampling.

Four Different Methods

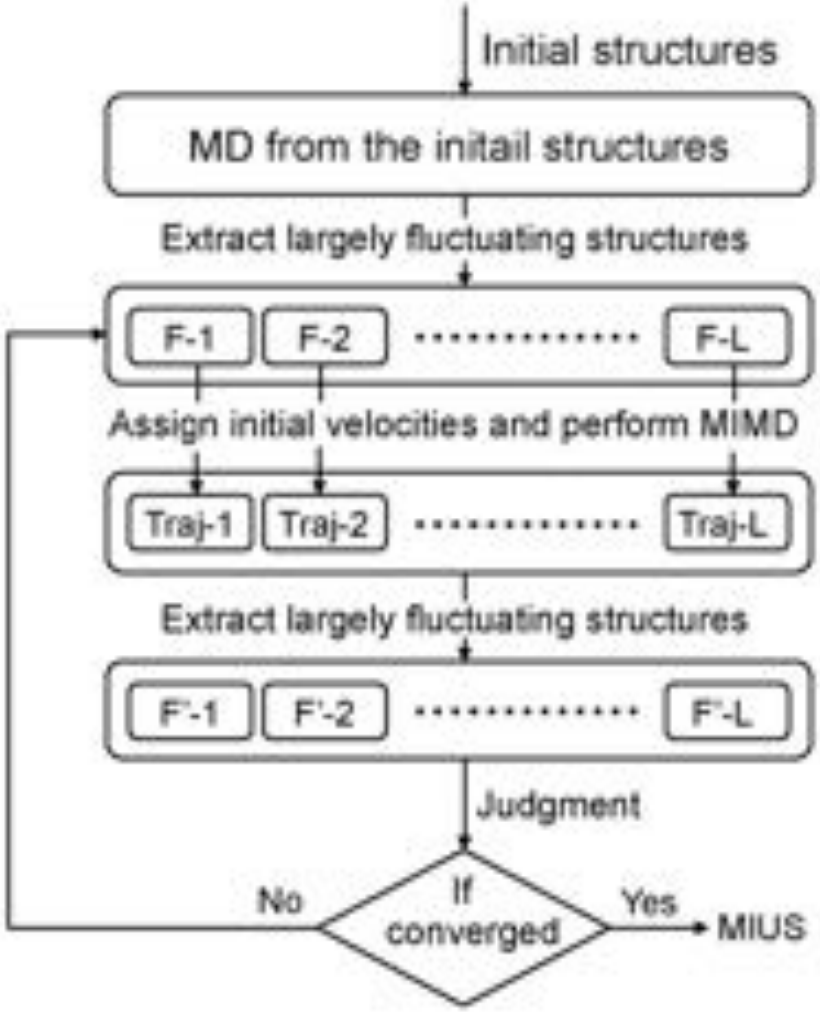
R. Harada, Y. Takano, T. Baba, **YS**, *Phys. Chem. Chem. Phys.* (***Feature Article***) **17**, 6155-6173 (2015)

Method	Structure of / Structures of both Initial condition	Dynamics	Measure
PaCS-MD	Reactant Product	Domain motion Induced fit Structural formation	RMSD, Bond, Angle, Radius of gyration
FFM	Reactant	Domain motion Ligand binding	Principal coordinates, Domain-Domain distance
OFLOOD	Reactant	Domain motion Protein folding Structural formation	Outlier
TBSA	Reactant	Protein folding	Energy

Fluctuation Flooding Method (FFM)

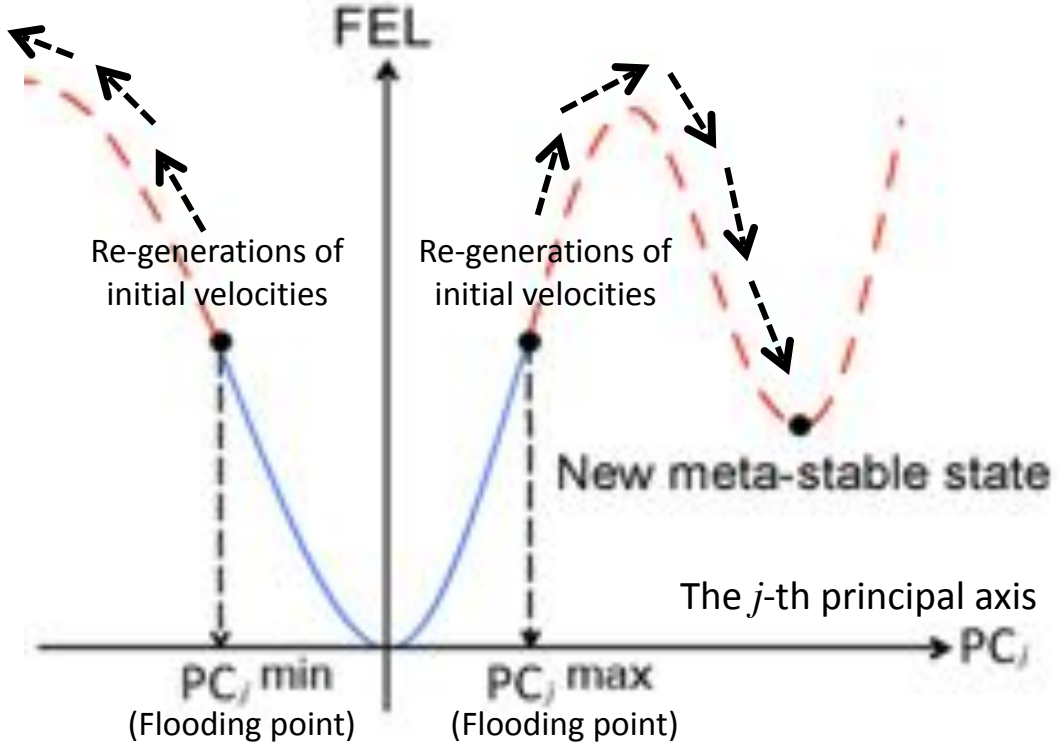
R. Harada, Y. Takano, **YS**, *J. Chem. Phys.* **140**, 125103 (2014)

◆Fluctuation Flooding Method (FFM)



◆Before a fluctuation flooding (the solid blue line)

◆After the fluctuation flooding (the broken red line)

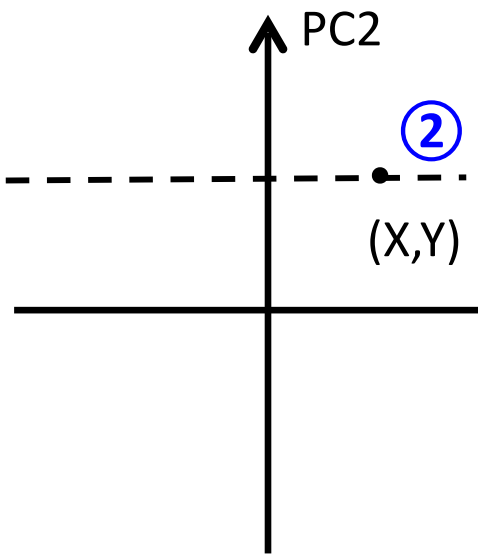


MIMD: Multiple Independent Molecular Dynamics simulations
MIUS: Multiple Independent Umbrella Sampling

Fluctuation Flooding for Principal Modes

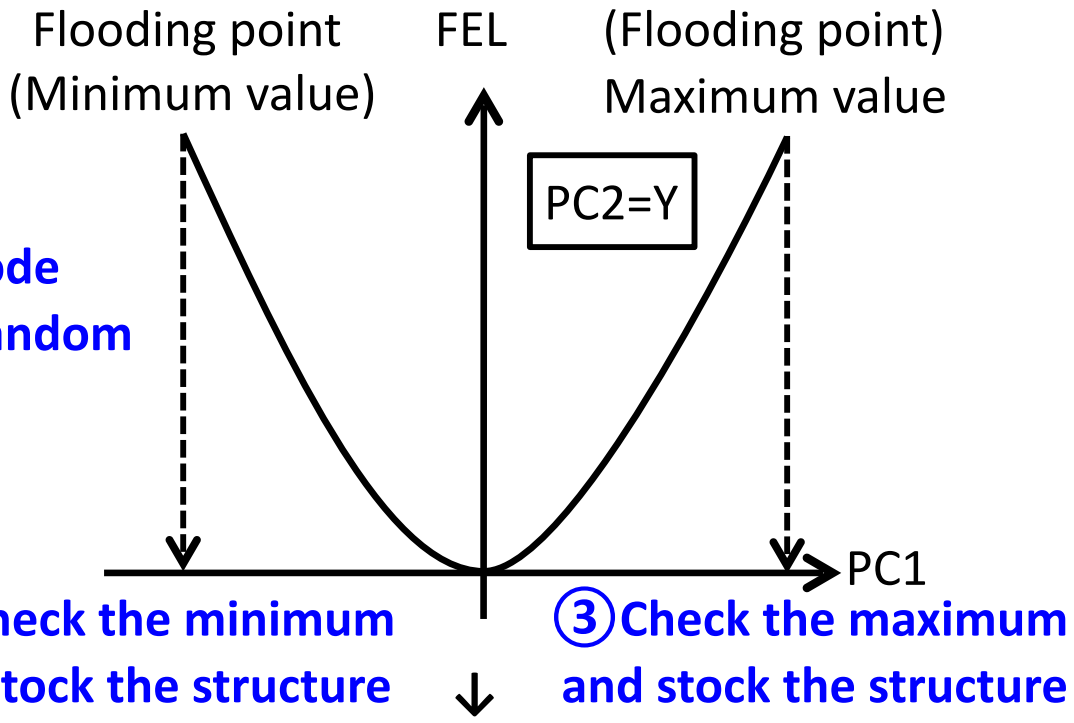
◆ **Consider the first and second principal modes to be enhanced in fluctuation flooding**

① **Select principal coordinates (PC1,PC2) = (X,Y) at random**



② **Select a principal mode to be enhanced at random**

→



③ **Check the minimum and stock the structure**

↓

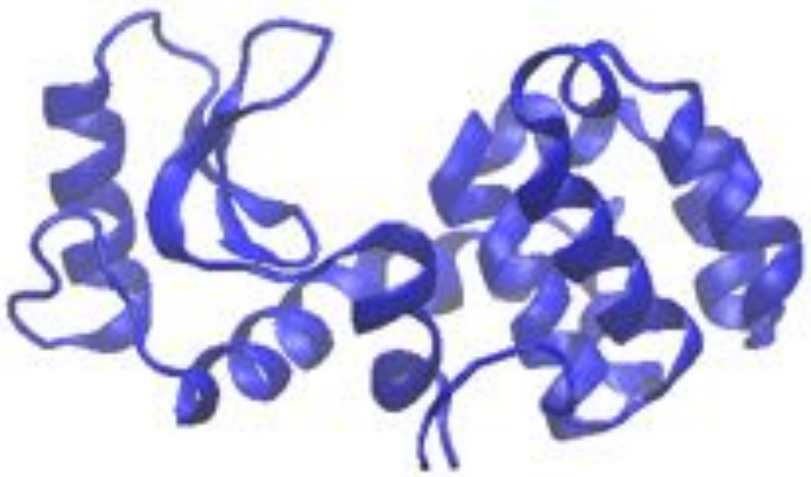
③ **Check the maximum and stock the structure**

④ **Perform MIMD starting from the flooding points via re-generations of initial velocities**
→Fluctuation Flooding from the minimum and maximum coordinates.

Application: T4 Lysozyme (T4L) in Explicit Water

◆T4-Lysozome (open structure)

164 residue (2643 atoms)



Open (150L, mutated back to WT)

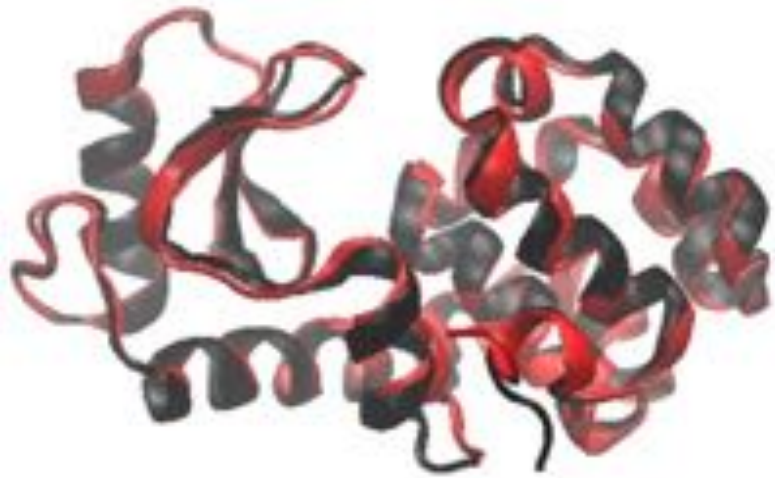
FFM

→

Conformational transition

◆T4-Lysozome (closed structure)

164 residue (2643 atoms)



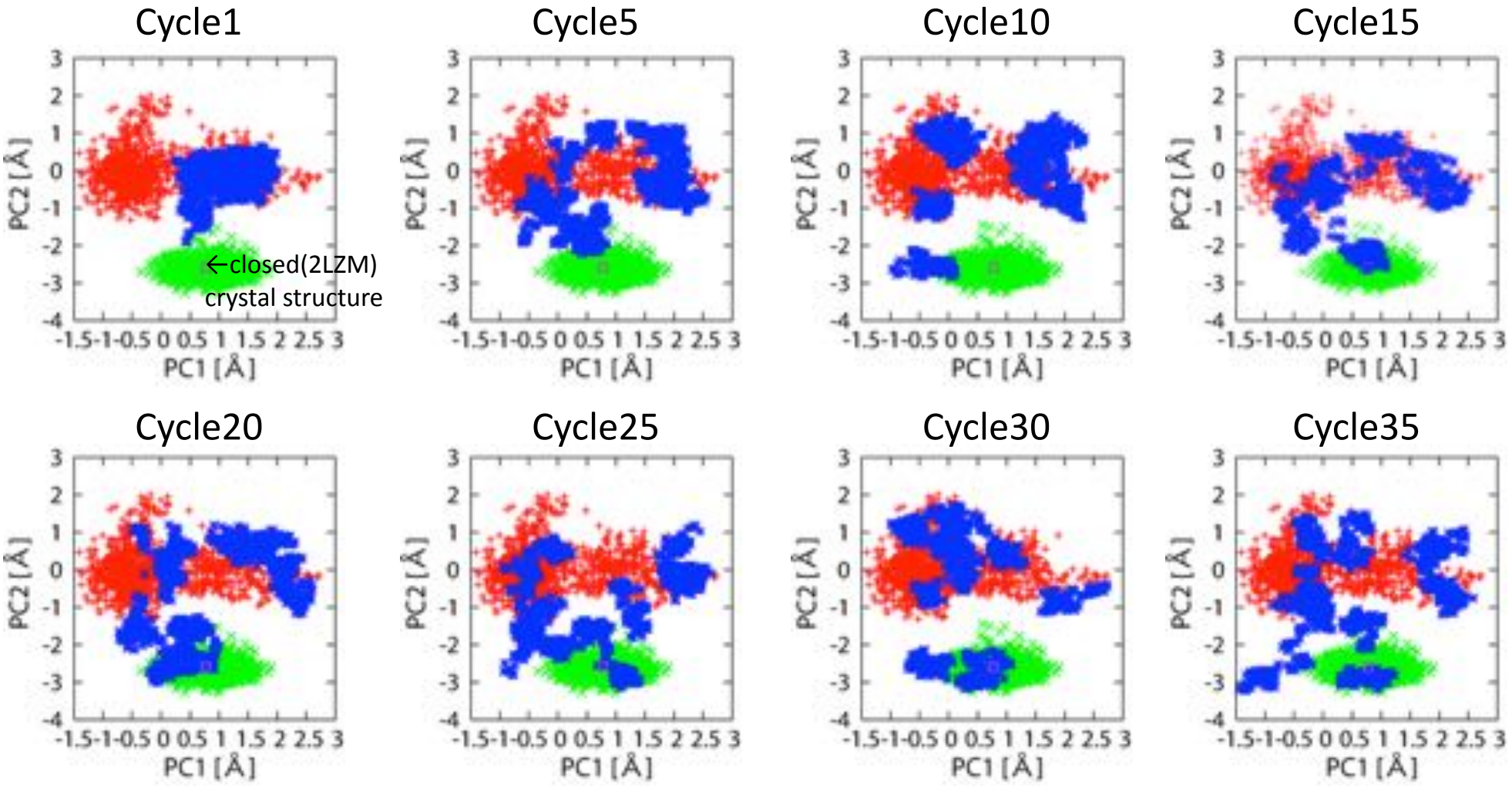
Closed (X-ray, 2LZM: Black)

Closed (**Predicted by FFM: Red**)

◆The minimum C_α RMSD ***0.76*** Å

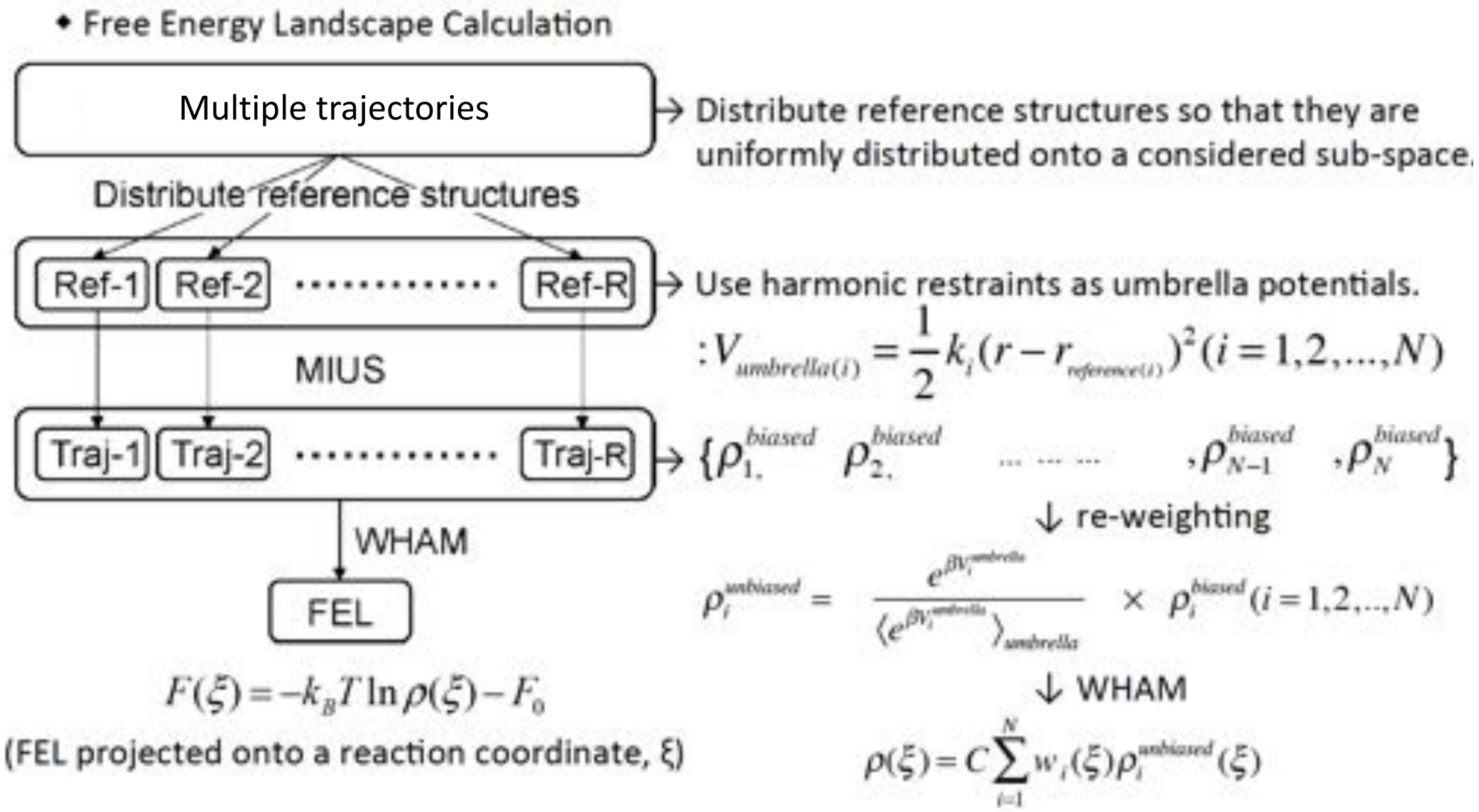
Parameters for MD simulations: AMBER PARM99SB, 2643 protein atoms, 8876 water molecule, 8 Cl^- under NVT (T = 300 K).

Projection onto PC1 and PC2 Space



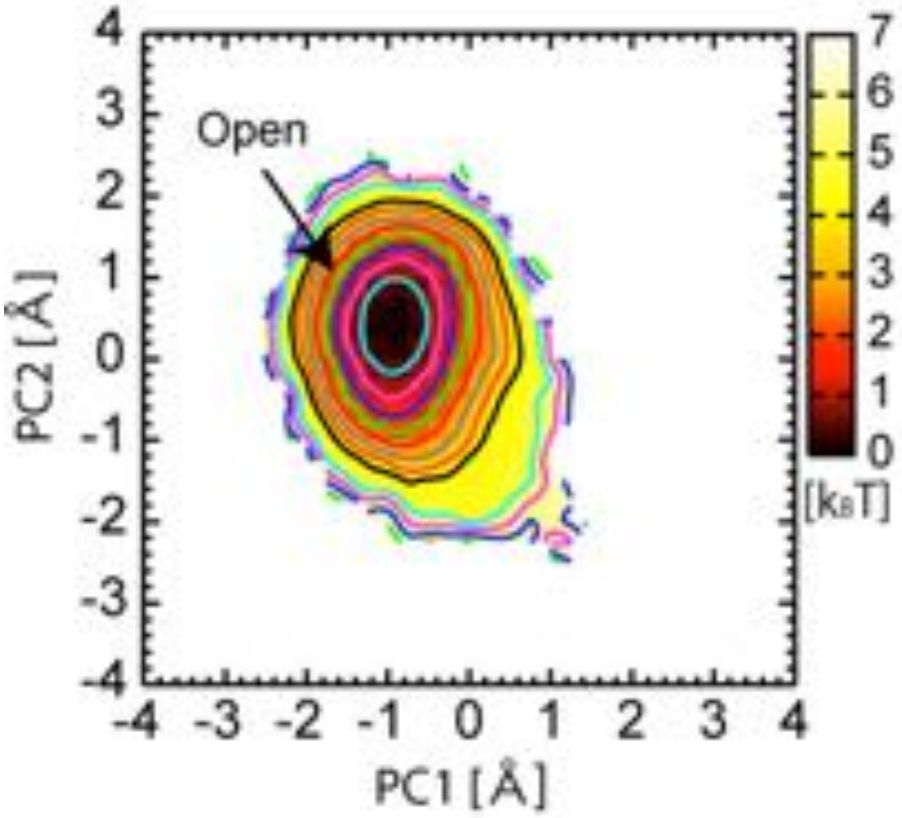
Red : 10 ns trajectory started from the open (150L, mutated back to WT)
Green: 10 ns trajectory started from the closed (2LZM)
Blue : 100 ps × 10 trajectories (for 10 candidates) obtained from FFM on each cycle.

Free Energy Calculation as Post-Processing Treatment

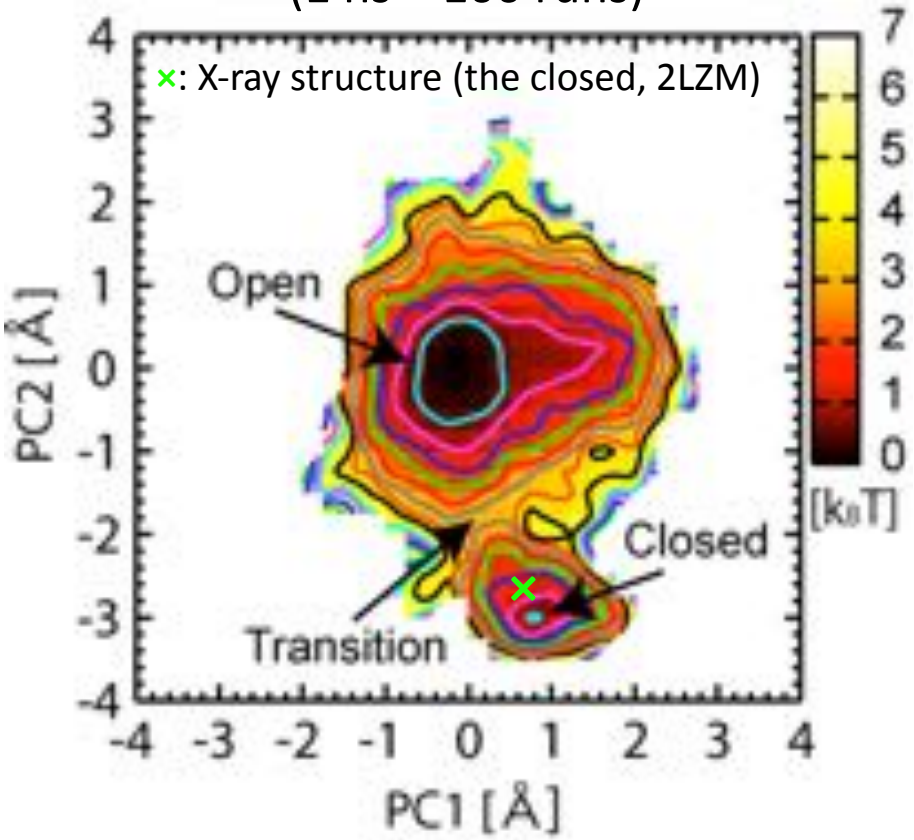


Free Energy Landscape of T4L (canonical MD vs. FFM)

◆Canonical long time (1 μs) MD started from the open



◆FEL calculated by FFM and MIUS (1 ns × 100 runs)

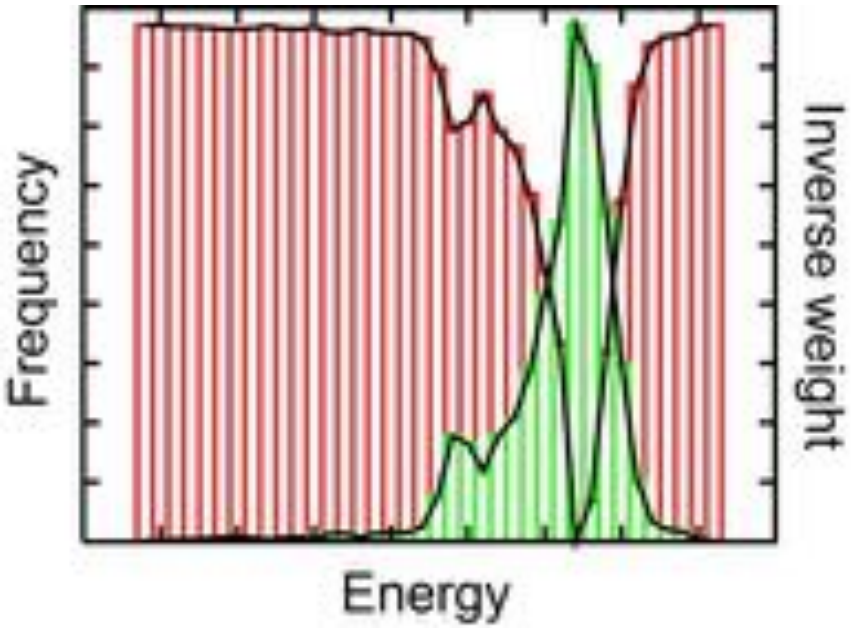


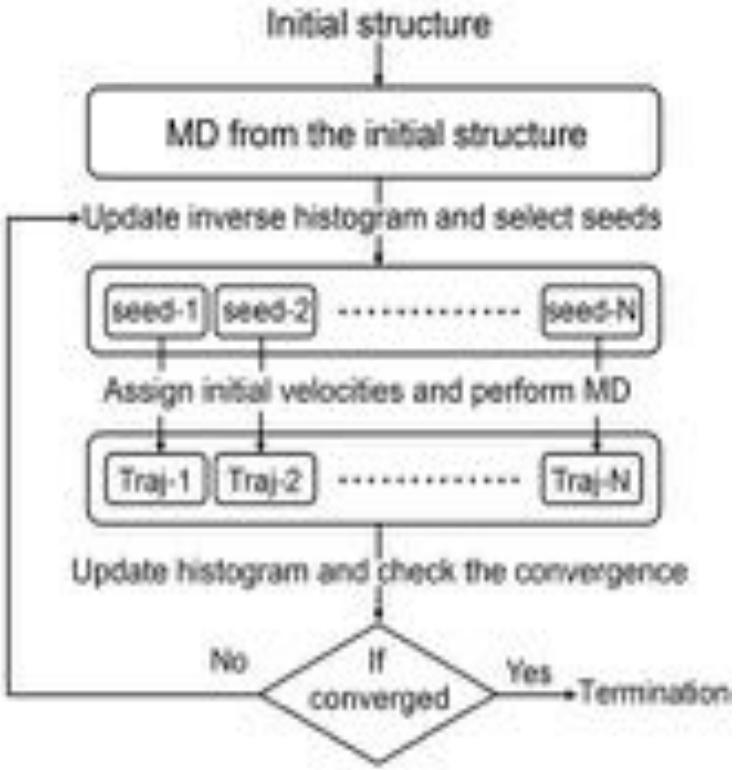
◆FFM could find the closed state as a local energy minimum, although 1-μs long time MD failed to find the closed structure!

TaBoo SeArch Algorithm (TBSA)

R. Harada, Y. Takano, **YS**, *J. Comp. Chem.* **36**, 763-772 (2015).

◆**Inverse weight for selecting seeds**

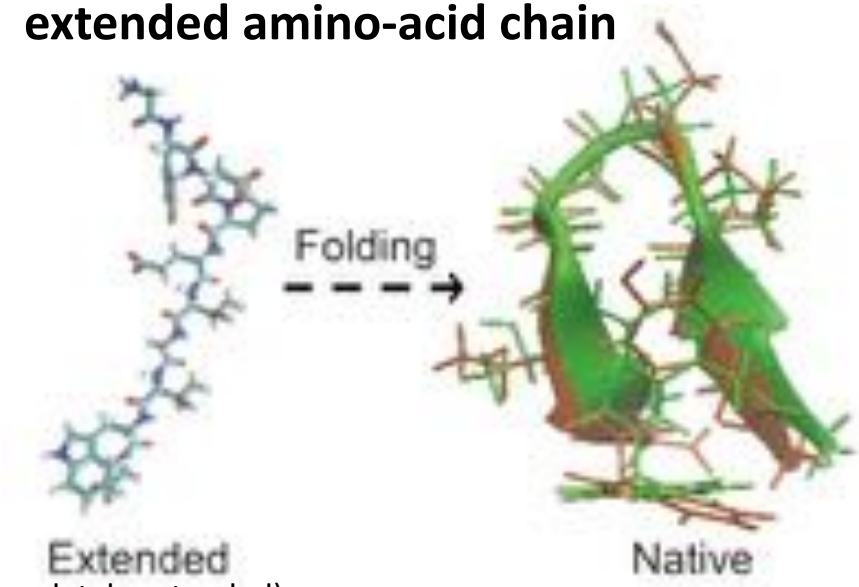




The schematic concept TBSA. The boxes colored in *green* correspond to *the original distribution in energy space* obtained from MD simulations. In contrast, the boxes colored in *red* correspond to *its inverse weights* for conformational sampling to explore energy spaces with low frequencies.

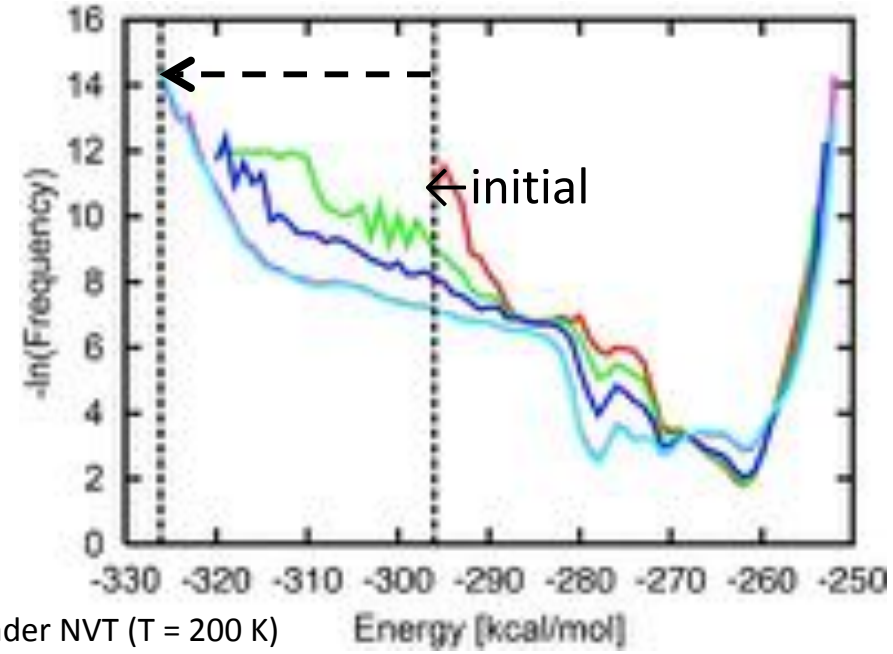
Application: Folding of Chignolin

◆**Folding of chignolin from a completely extended amino-acid chain**



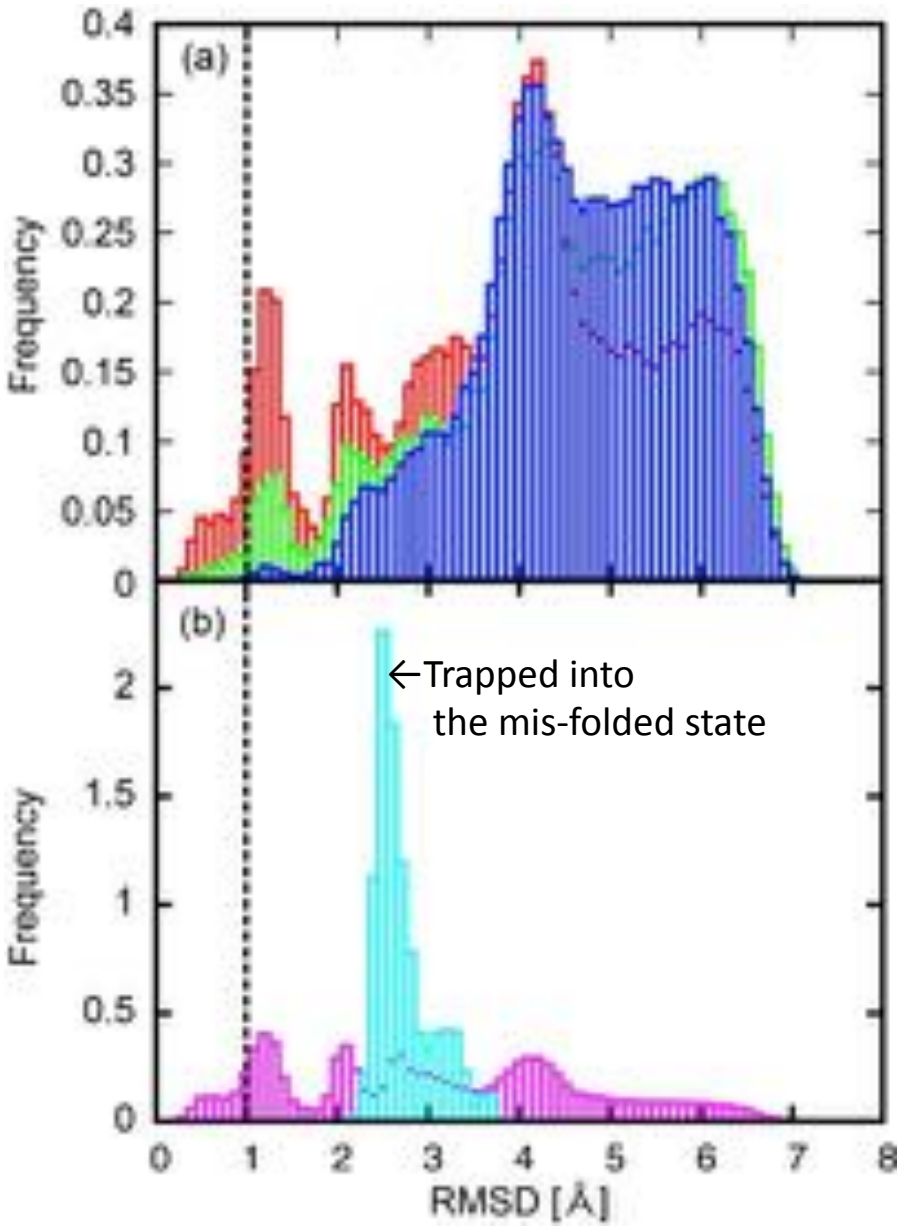
❖Simulations done by generalized born model (igb = 5 in AMBER 11) under NVT (T = 200 K)

◆**Convergences of distributions of energy**



The convergences of distributions on energy spaces every cycle in Taboo search algorithm. The lines colored in red, green, and blue correspond the distributions of the first , the second, the third cycles, respectively. The lines colored in cyan and magenta correspond to the ones of last two cycles, the 9-th and the 10-th. The lowest energy values in the first and the last cycles are highlighted by the broken lines.

Distributions of RMSDs

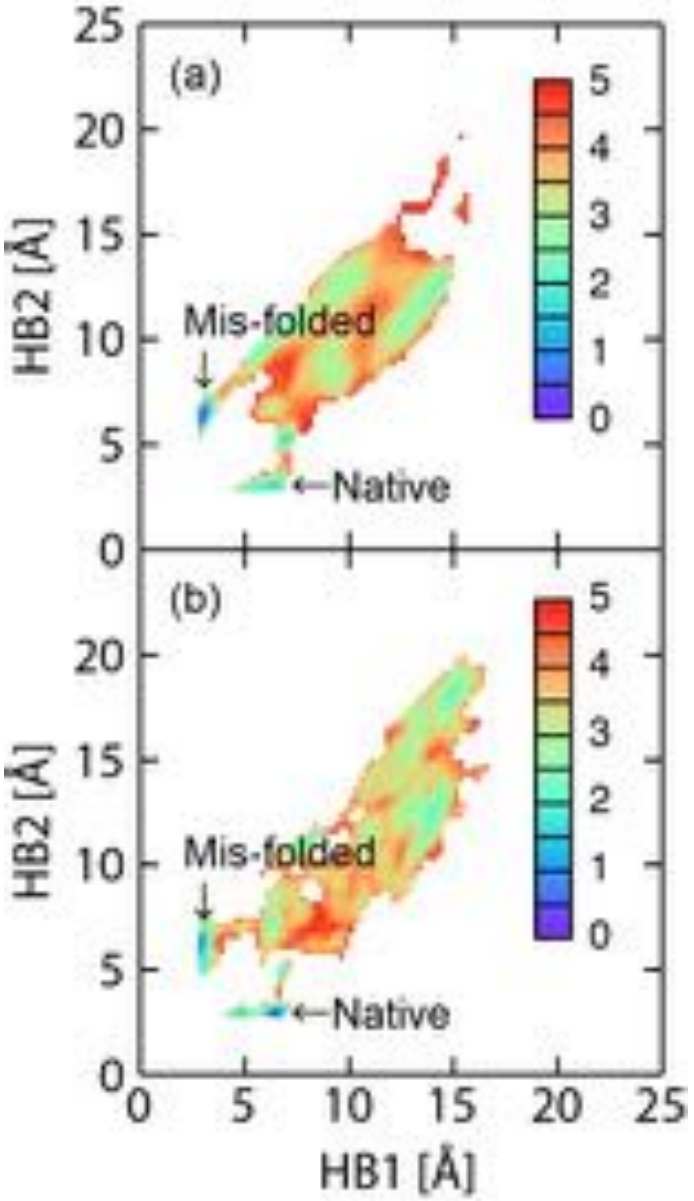


(a) The distributions on C_α RMSD in the first three cycles. The boxes colored in blue, green, and red correspond to the distributions of the first, the second, the third cycles, respectively. (b) The comparisons with a brute force 1-μs CMD simulation. The boxes colored in magenta and cyan correspond to the distribution of the brute force CMD and the last cycle (the 10-th cycle) of the taboo search algorithm. The dashed line indicates a criterion for folding into the native structure (C_α RMSD < 1.0 Å)

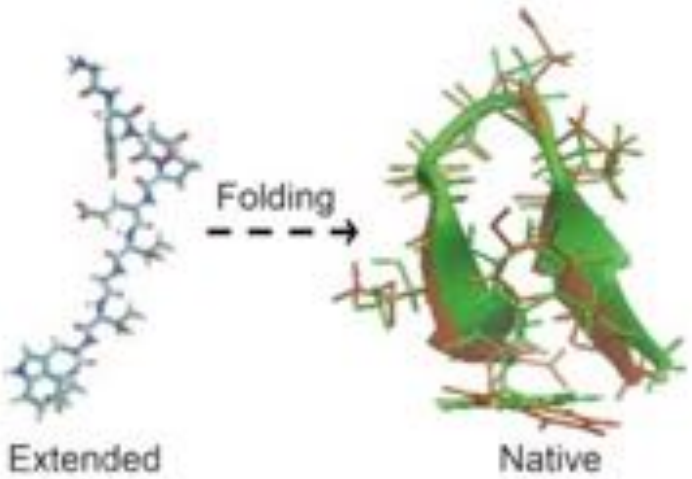
◆Minimum time to sample native structure

TBSA Trial Index	Minimum C_α / all-atom RMSD [Å]	Minimum simulation time [ns]
CH1	0.14 / 1.79	16.1
CH2	0.25 / 1.91	59.0
CH3	0.23 / 1.79	49.6
CH4	0.17 / 1.69	37.5
CH5	0.21 / 1.76	45.0
CH6	0.86 / 2.73	78.0
CH7	0.37 / 2.28	235.9
CH8	0.22 / 1.52	99.1
CH9	0.20 / 1.82	15.8
CH10	0.16 / 1.63	13.1

Folding Free Energy Landscape of Chignolin



(a) Folding landscape of chignolin in implicit solvent calculated by MIUS. References for the MIUS were randomly selected from snapshots obtained by TBSA (200 references). (b) Folding landscape of chignolin in explicit solvent. The coarse-grained landscape in implicit solvent was refined through the multi-scale free energy landscape method (MSFEL: Chem. Phys. Let., **503**, 145, (2011)) with the explicit solvent model.

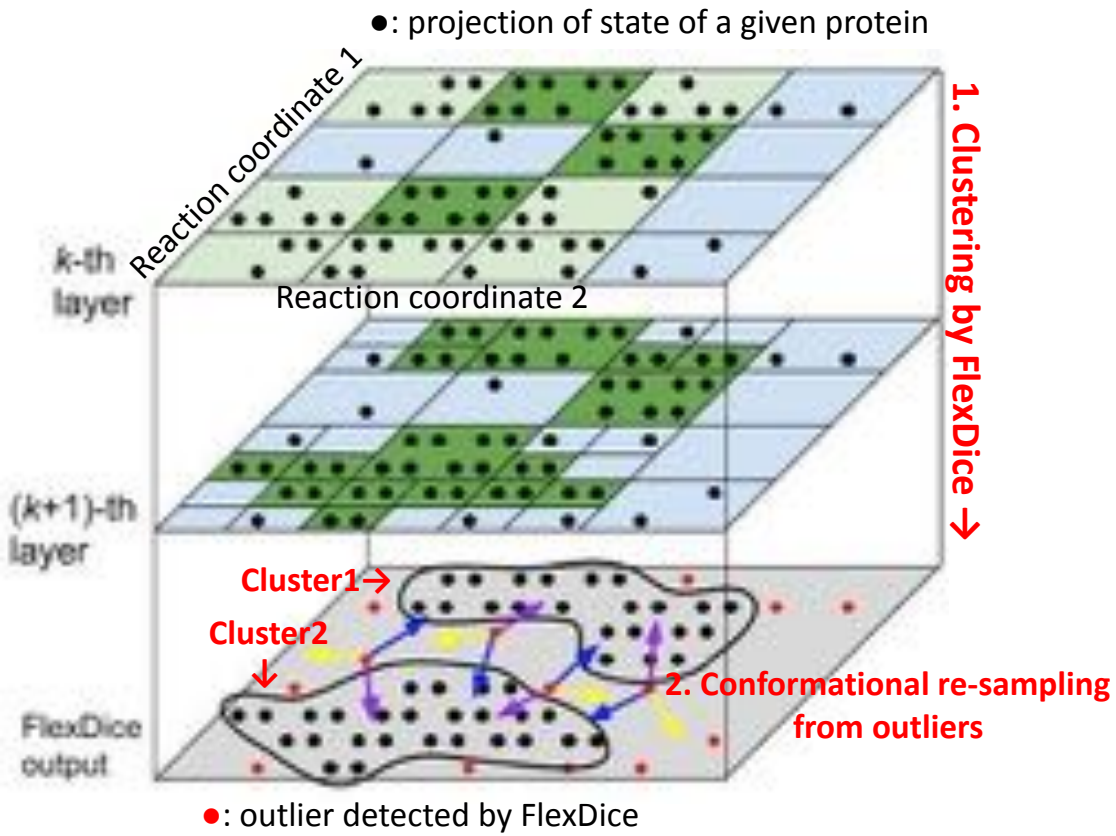


Native structure (green) and snapshot sampled by TBSA (brown).
TBSA sampled the native structure with *0.22 Å in C_α RMSD* from the native.

Outlier FLOODing Method (OFLOOD)

R. Harada, T. Nakamura, Y. Takano, **YS**, *J. Comp. Chem.* **36**, 97 (2015).

◆**Detections of outliers from distributions**



When we perform projection of trajectories onto a set of reaction coordinates, there exist several clusters on it.

The outliers are defined as complements to clusters and are located nearby the edges of clusters.

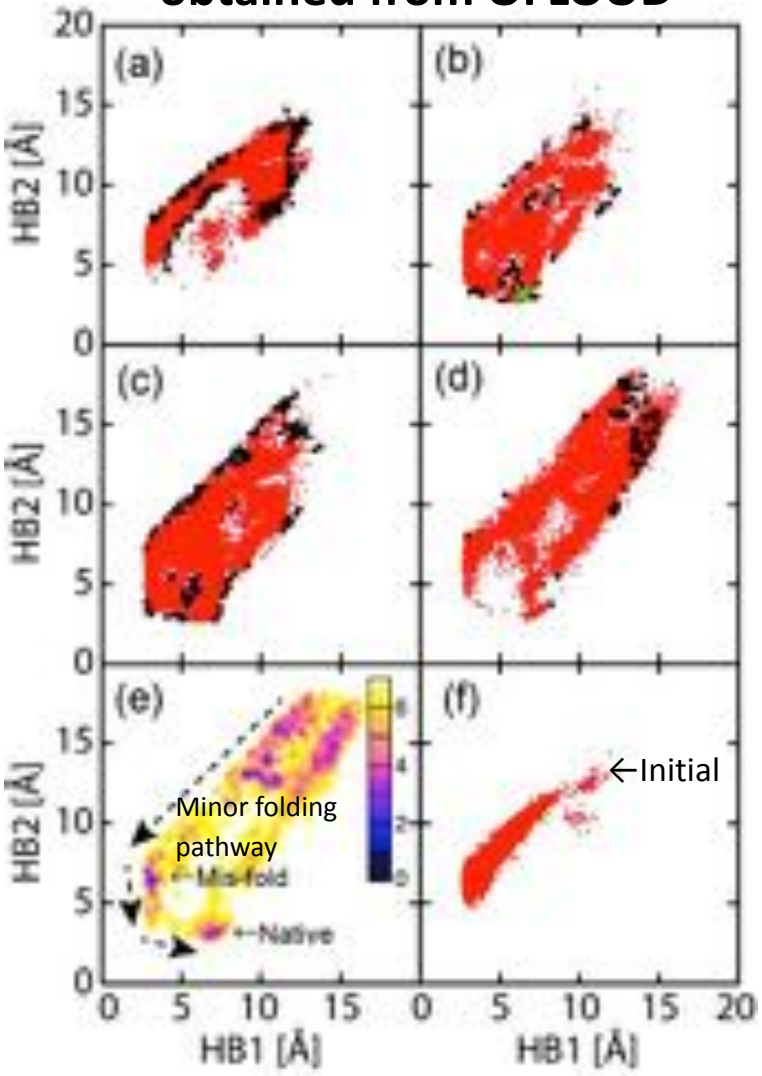
Since some of outliers are in between two clusters (one is known and other is unknown), if short-time MD starts from outliers, MD might eventually find a new cluster.

1. **Detections of outliers by FlexDice, a hierarchical clustering algorithm.**
2. **Conformational re-sampling of the outliers via re-starting MD (re-generating initial velocities**

❖**Repeating the above schemes updating outliers.**

Demonstration: Folding Pathway of Chignolin

◆**Projections of trajectories obtained from OFLOOD**



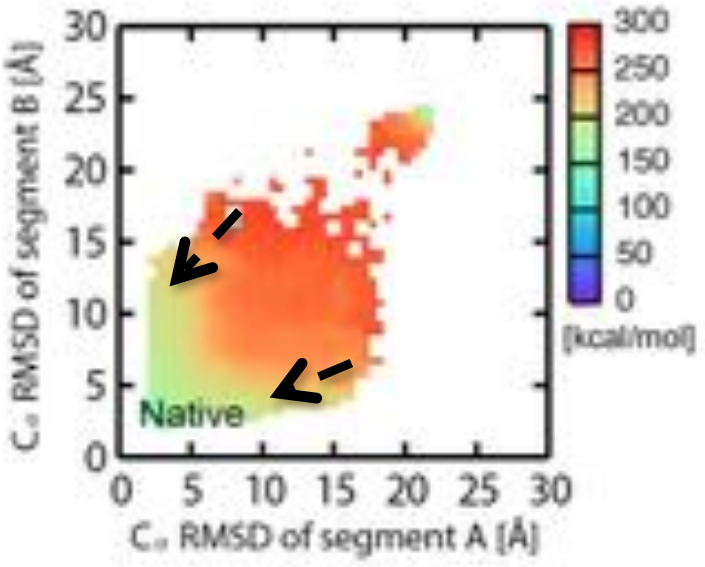
(a-d) Projections of the trajectories (red) after the conformational re-sampling from the outliers (black) at the 2nd, 4th, 6th, and 8th cycles, respectively. (Totally, 10 cycles) ***The cross in (b)*** represents the native.

(e) The population (the minus common logarithm of frequencies) of ***the reactive trajectories jointing the snapshots in the initial and final cycles***. The minor folding pathway extracted by OFLOOD is indicated by the dashed arrows.

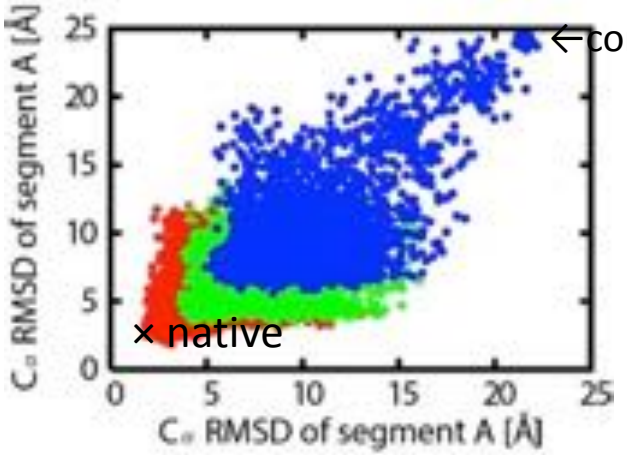
(f) The projections of the trajectories of ***the brute-force (1-μs) CMD simulation***, ***initiated from a completely stretched structure***.

1BBD (46 residues, 725 atoms)

Averaged potential energy surface obtained by OFLOOD



Projections of trajectories obtained from OFLOOD



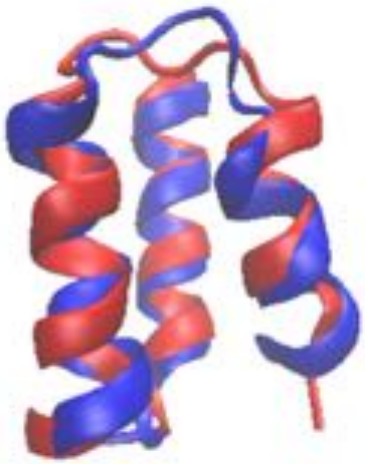
(Blue: cycle0, Green: cycle11, Red: cycle30)

The minimum computational time to sample the native : *185.9 ns*

cf. 3.9 μs by ANTON for α3D

The minimum C_α RMSD from native structure : *2.10 Å*

cf. 3.3 Å by ANTON for α3D



Blue: X-ray structure

Red: The snapshot with the minimum C_α RMSD sampled by OFLOOD

❖Simulations done by Generalized Born model (igb=5 in AMBER 11)
Initial structure: Completely extended amino-acid chain.

Summary

- ***Conformational re-sampling of seeds*** (***snapshots with high potentials to induce structural transitions***) is effective for extracting biologically rare events of proteins (such as **Domain motion, Protein folding, and Induced-fit (not shown today) processes**).

- The most important thing is ***how to select the seeds based on appropriate measure(s)***.

- A ***distributed computing method*** might be a powerful approach, if massively computational resources are available.