Markov-state modeling of biomolecular systems pt. I

Toni Giorgino
National Research Council of Italy
toni.giorgino@cnr.it
www.giorginolab.it

@giorginolab

Projects available!

Master in Bioinformatics for Health Sciences
UPF Barcelona, 6 May 2021

Introduction

- The aim of this class is to provide a practical *overview* of Markov state models in computational structural biology
- MSM emerging because:
  - reconstruct *kinetic* information*, including state transition *networks*, from simulated trajectories
  - start from ***unbiased simulations*** (no *a priori* reaction coordinate hypothesis necessary)
  - microsecond-scale (high-throughput) trajectory data are becoming accessible (e.g., with GPUs)
- Success cases: *ab initio* folding, drug binding, peptide binding, …

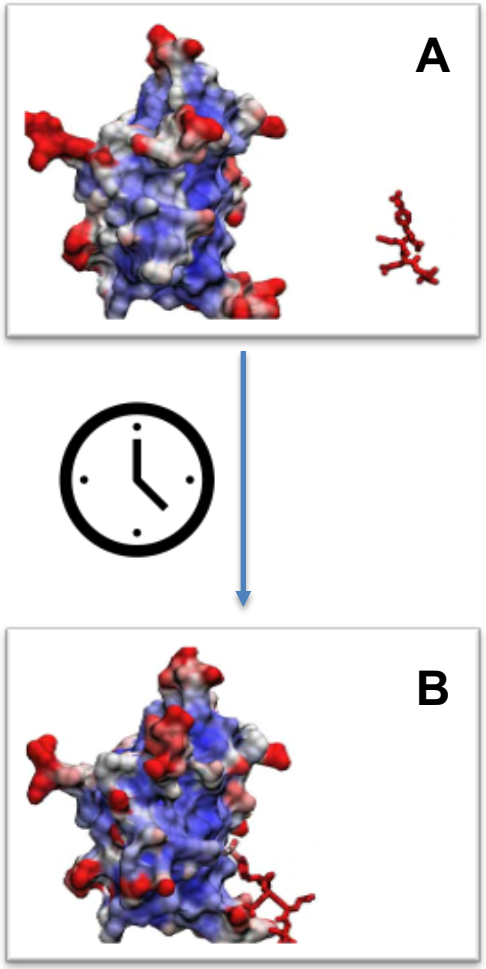
* as well as the corresponding structures

Motivation

- Can we use an *ensemble* of simulations to estimate dynamic behaviours which happen on timescales longer than each of the observed trajectories?
- In other words, can we leverage several "short-sighted" (non-equilibrium) observations to *extrapolate* long-time behaviour?

The general idea

Unbiased sampling



System evolves in time*.

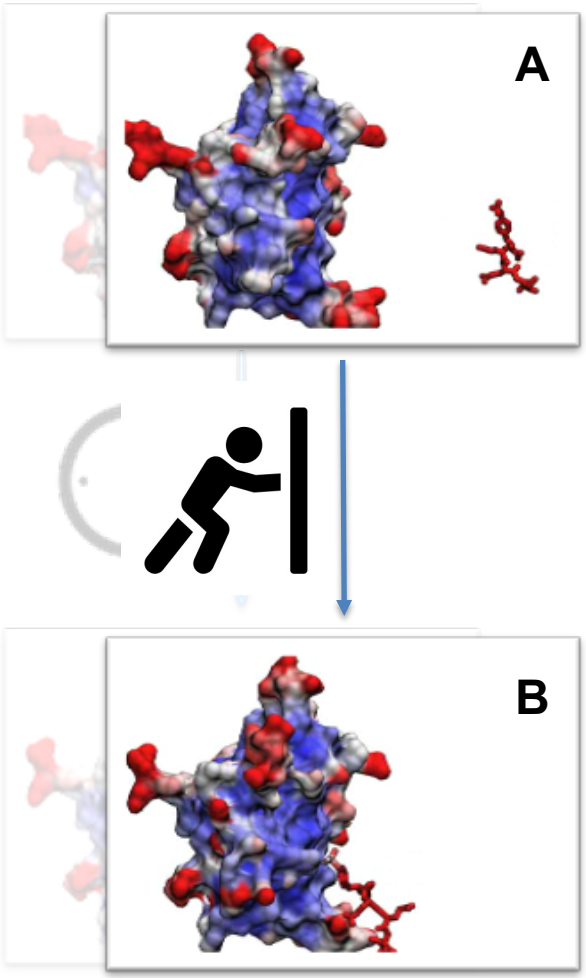
We do not "guide" the system with biases in any particular direction.

Pro: Kinetics are "true"

Con: Slow. Observations (=sufficient sampling) may well be unfeasible.

* Under the influence of internal end external forces designed to approximate "real" ones.

Biased sampling



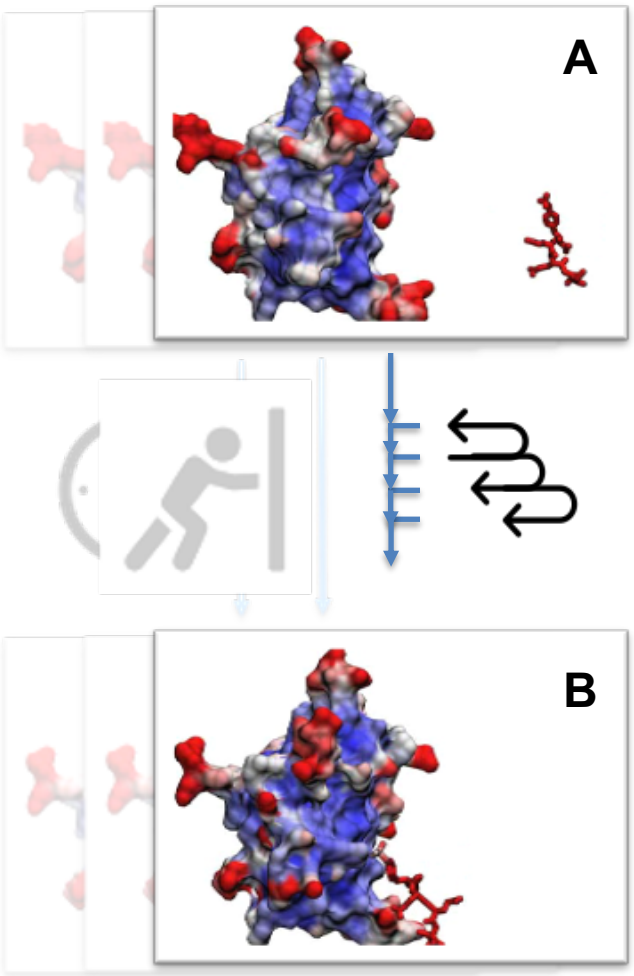
Various strategies* exist to "push" the system *towards* desired states

Pros: Sampling is accelerated..
Can recover ΔG.

Cons: Biasing "direction" must be set, often not trivial.
Kinetics (in general) are altered.

* Examples: metadynamics, steered MD, …

Markov-state modeling



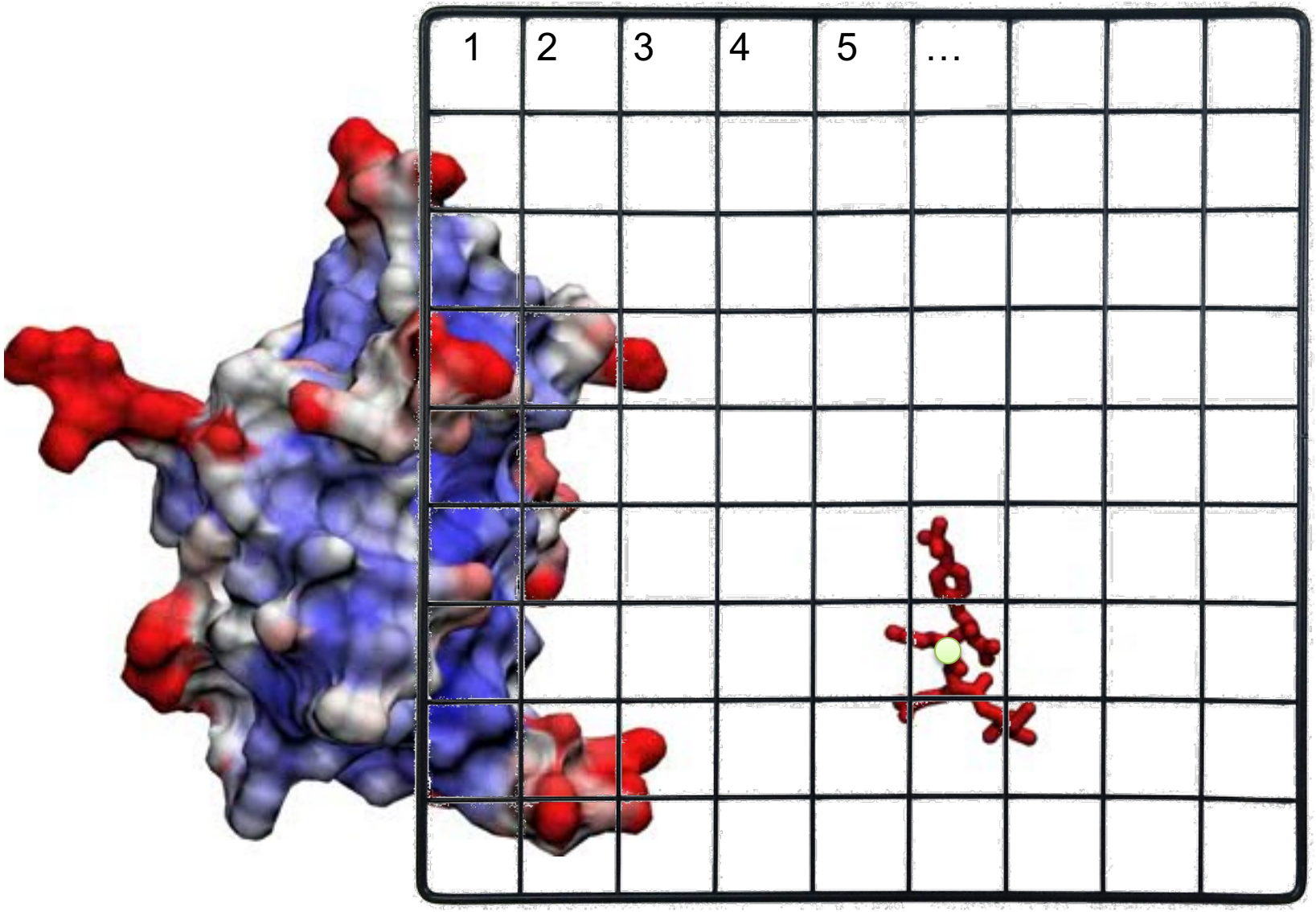
Decompose a system in multiple *states*. Use shorter trajectories to compose a statistical* picture.

Pros: True kinetics.
No need to pre-set a reaction coordinate.

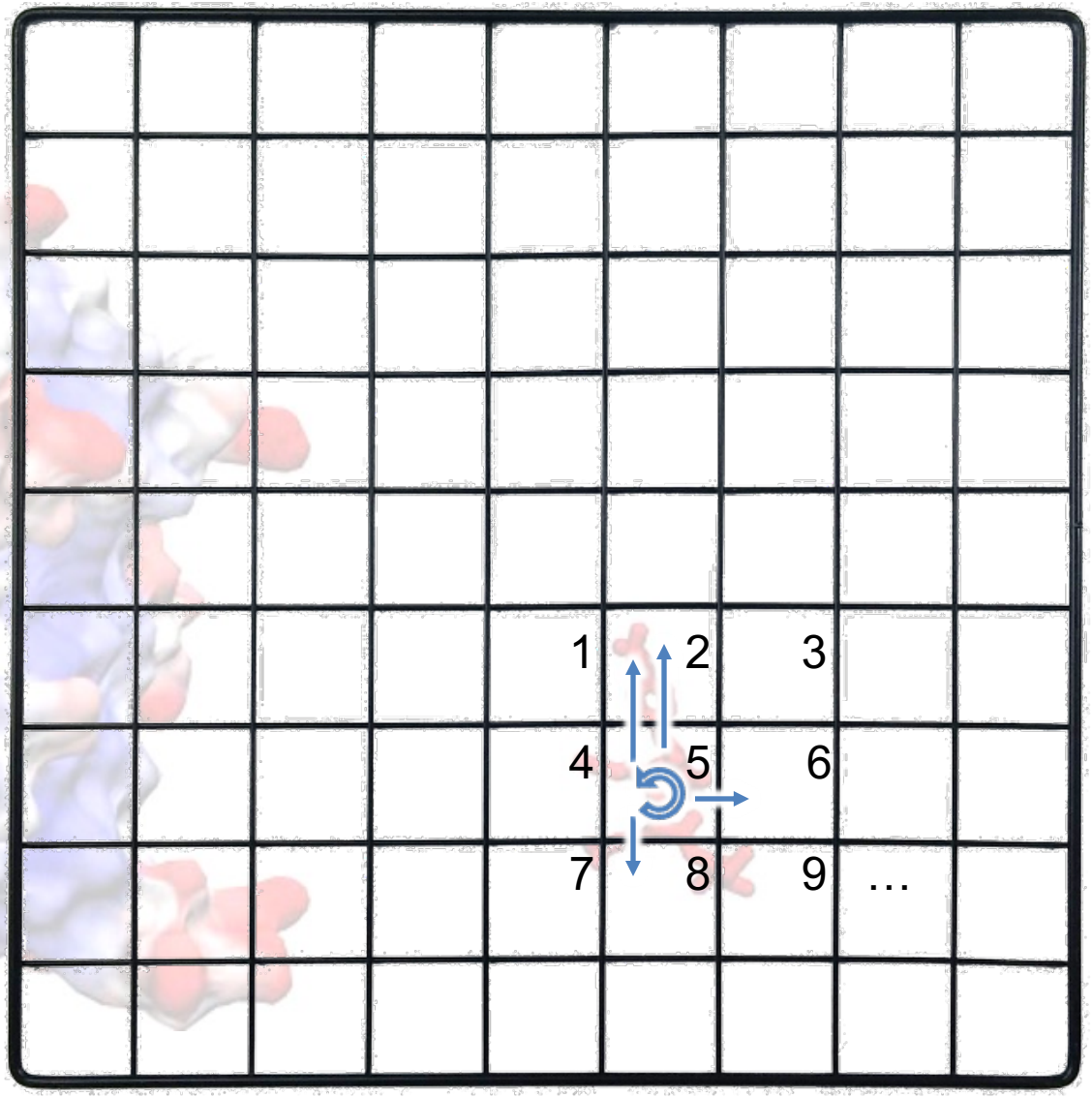
Cons: Non obvious (but software helps).
Sampling still needed for convergence.

* We'll see Markov-based formalism, but others exist, such as *milestoning*

Decomposing the state space


1
2
3
4
5
…

Count the state transitions made within a time interval (τ)



$T_{56}=1$

$T_{52}=2$

$T_{58}=1$

$T_{55}=1$

Total: $N_5=5$

$T_{56}=1$ $p_{56} = 1/5 = 0.2$

$T_{52}=2$ $p_{52} = 2/5 = 0.4$

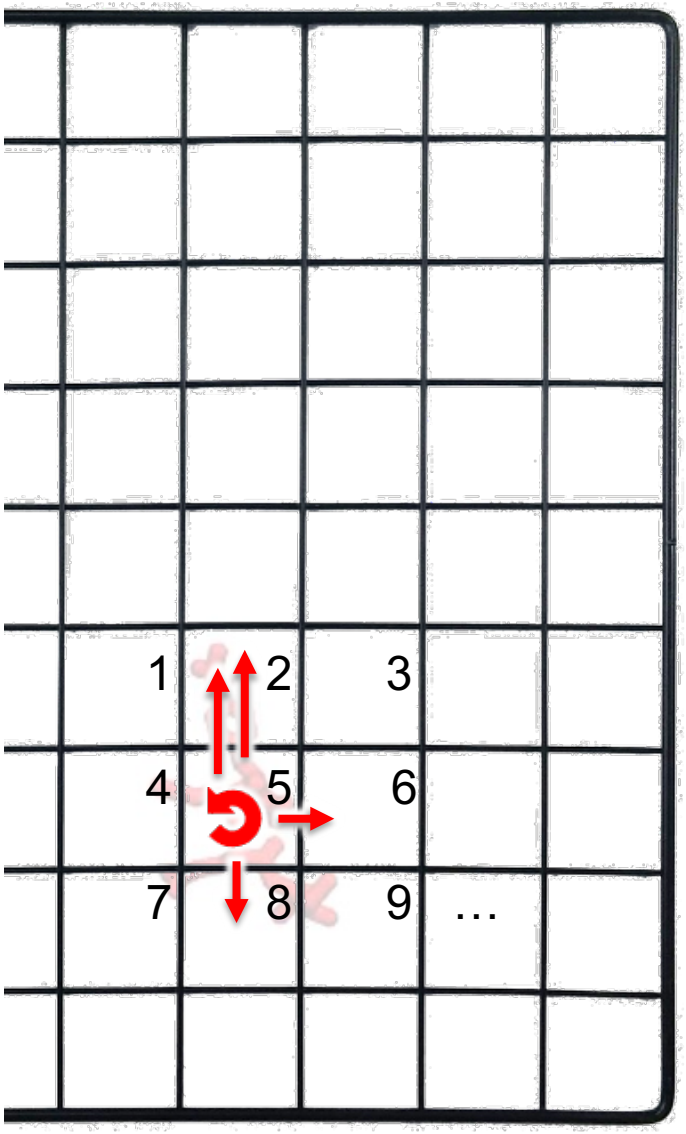
$T_{58}=1$ $p_{58} = 1/5 = 0.2$

$T_{55}=1$ $p_{55} = 1/5 = 0.2$

$N_5=5$ $p_{5*} = \Sigma_i p_{5i} = 1$

	1	2	…	5	6	7	8	…
5		0.4		0.2	0.2		0.2	

Row 5 of the
transition probability matrix P

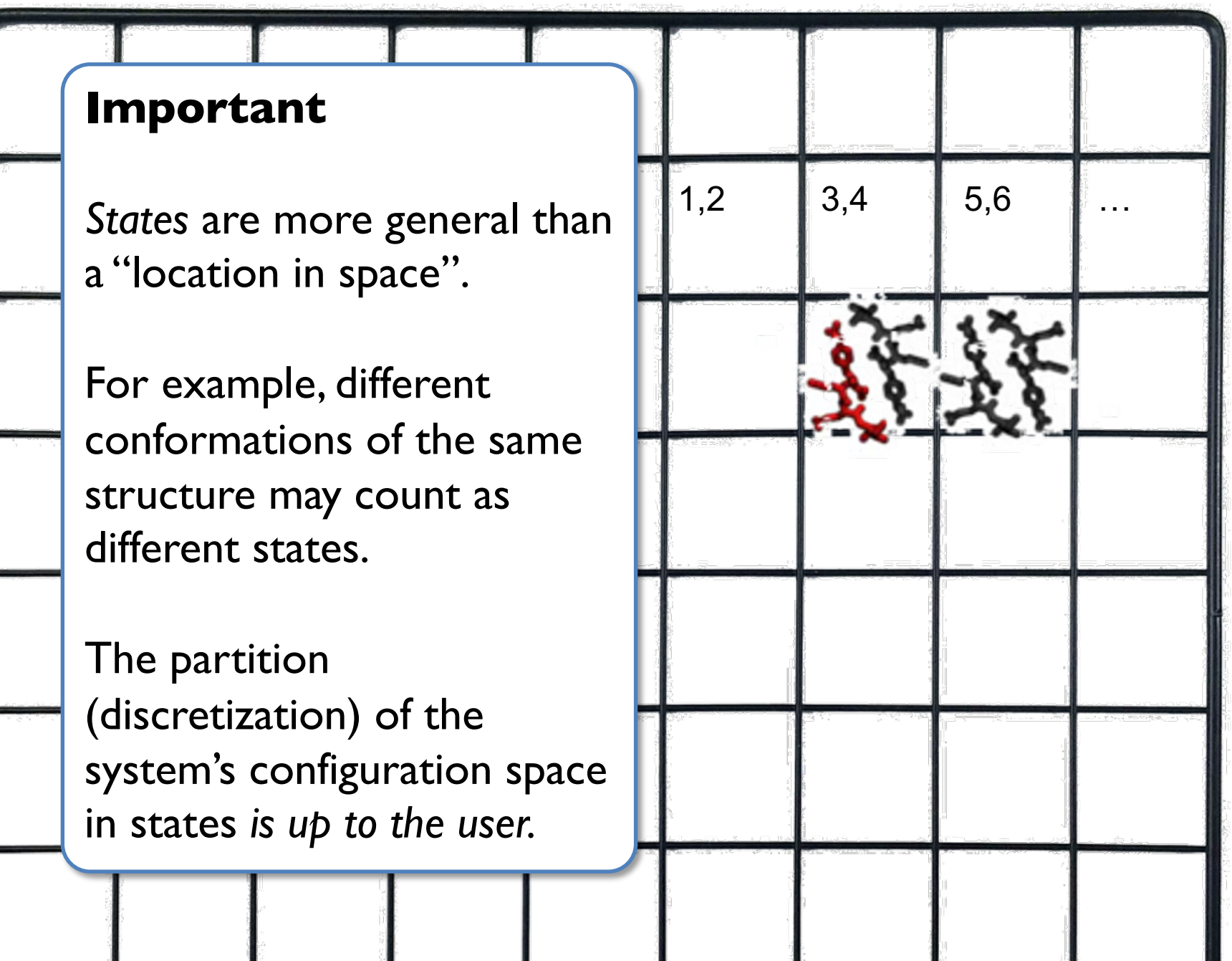


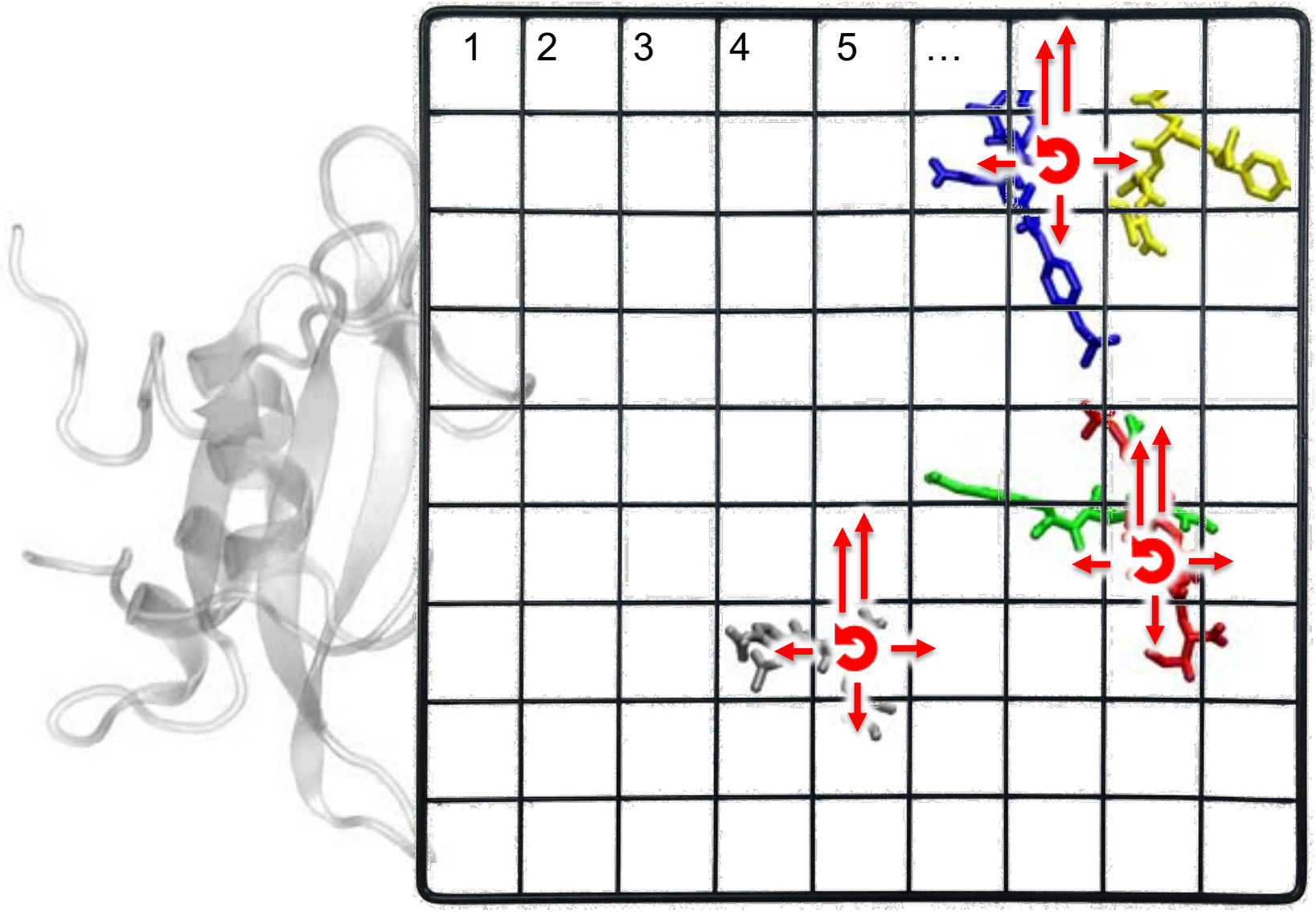
Important

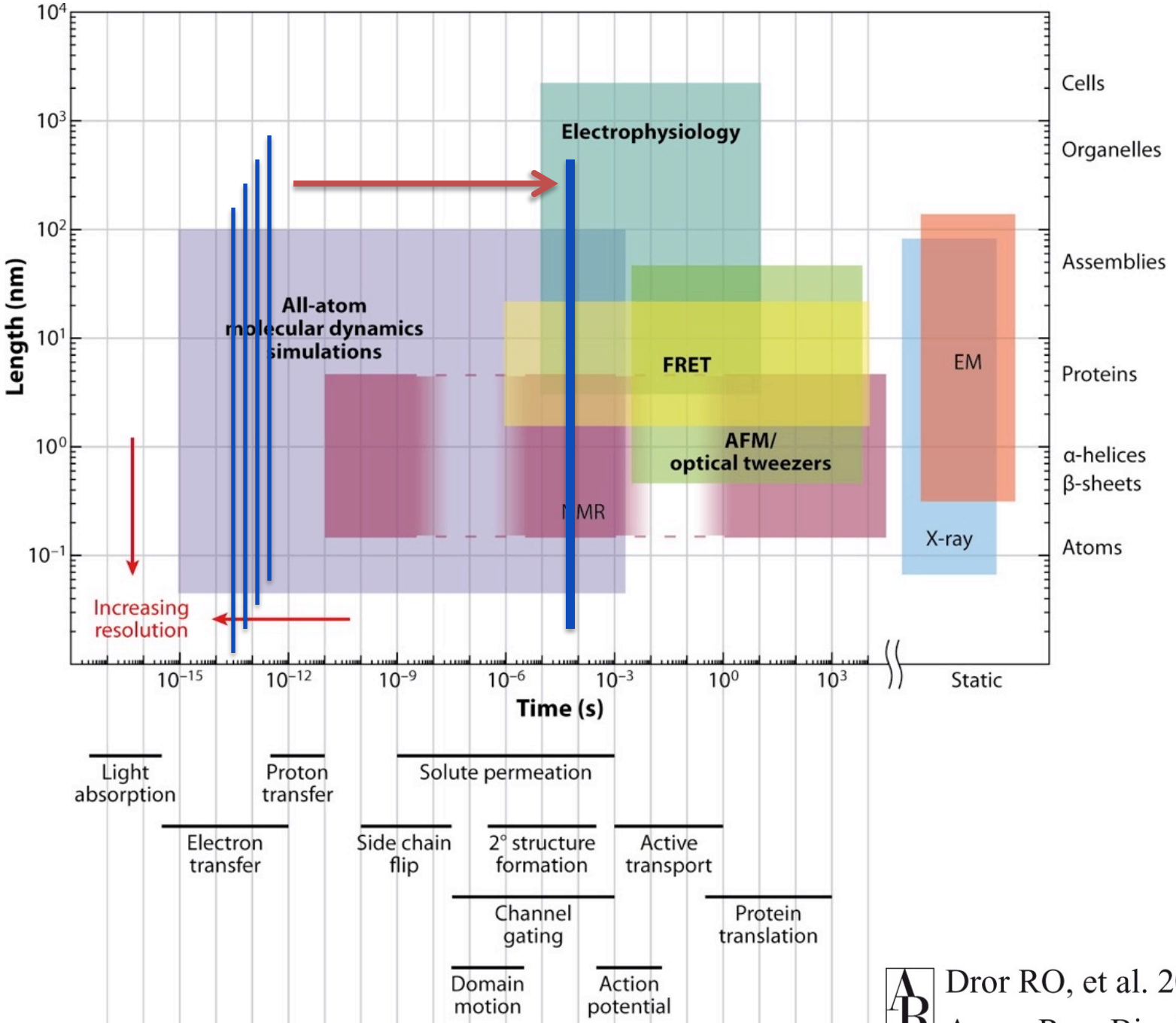
States are more general than a "location in space".

For example, different conformations of the same structure may count as different states.

The partition (discretization) of the system's configuration space in states *is up to the user.*




1
2
3
4
5
…



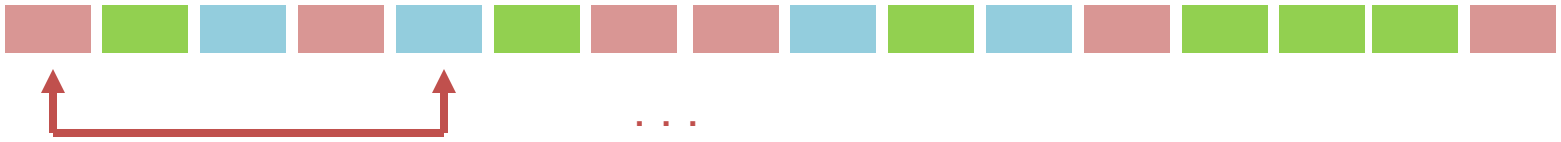
Dror RO, et al. 2012.
Annu. Rev. Biophys. 41:429–52

Learning transition matrices from trajectories

Markov models

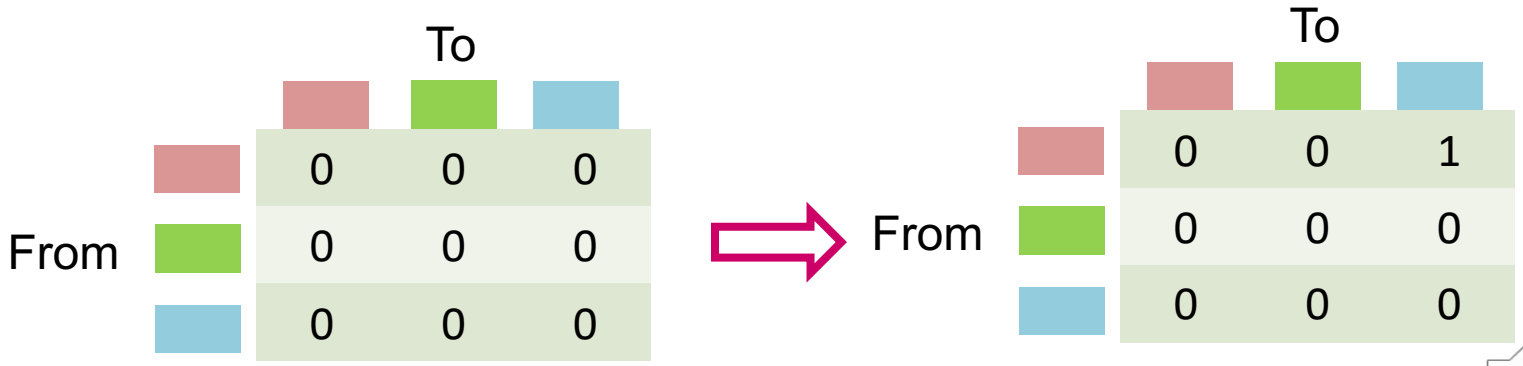
Total sampled time (e.g. 100 μs)

Define the *discrete* state of the system in *discrete time* (e.g. via reaction coordinates)



Lag time τ (here: 4 time units)

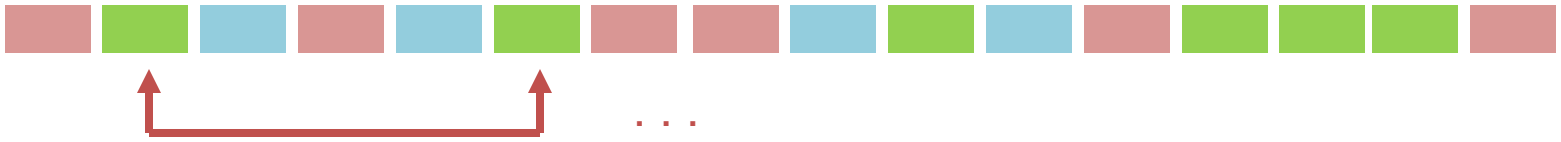
Compute the transition probability matrix sliding a window of lag time τ



Markov models

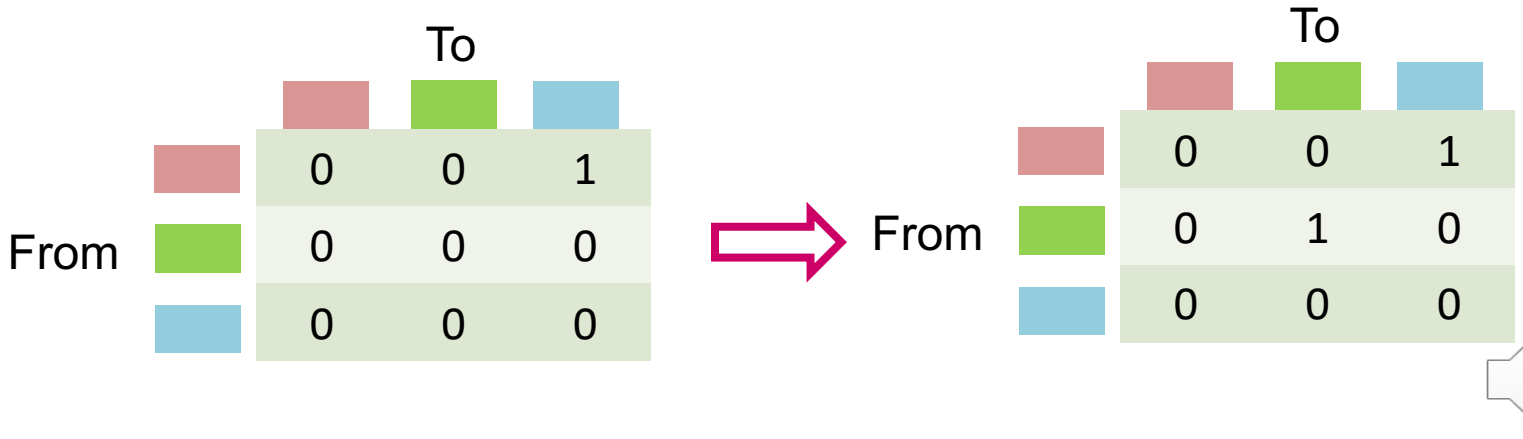
Total sampled time (e.g. 100 μs)

Define the *discrete* state of the system in *discrete time* (e.g. via reaction coordinates)



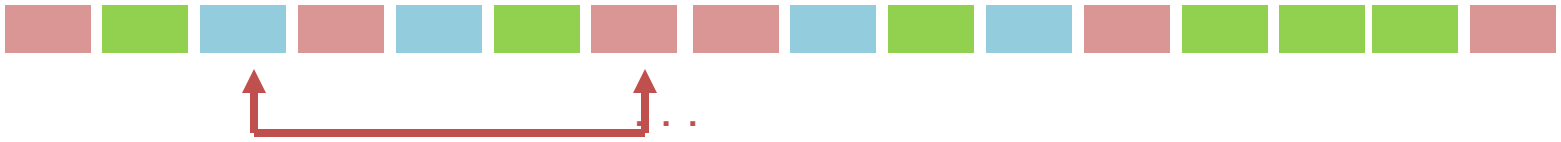
Lag time τ (here: 4 time units)

Compute the transition probability matrix sliding a window of lag time τ



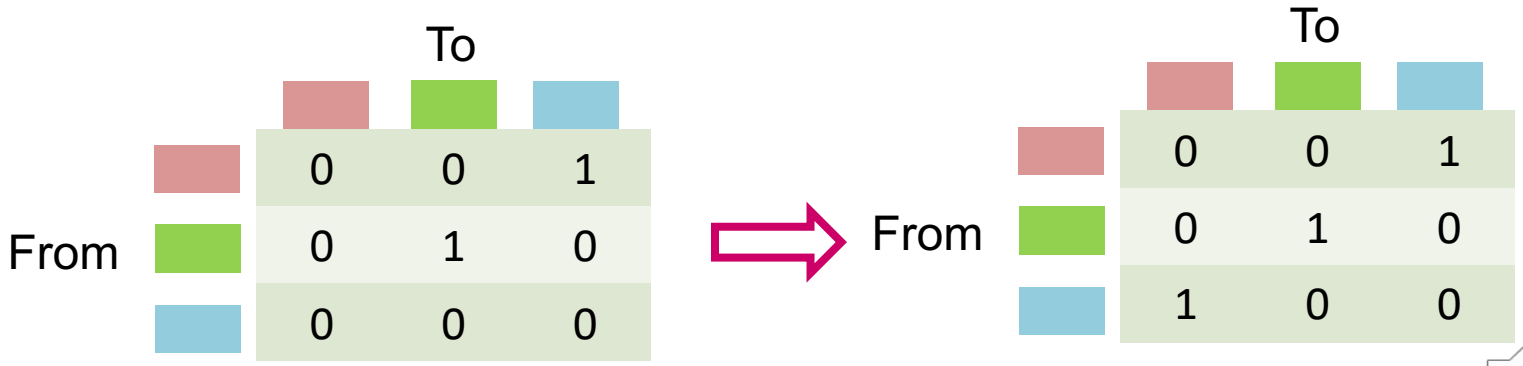
Markov models

Define the *discrete* state of the system in *discrete time* (e.g. via reaction coordinates)



Lag time τ (here: 4 time units)

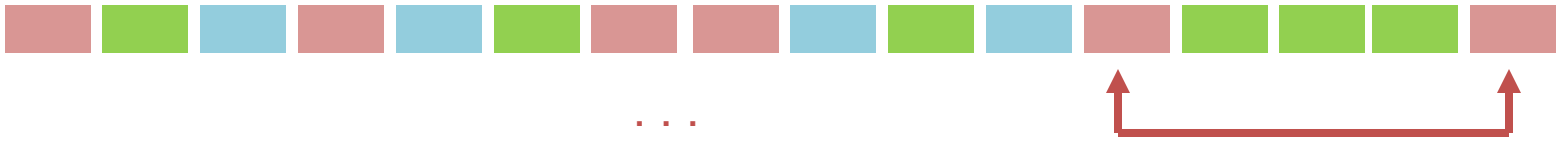
Compute the transition probability matrix sliding a window of lag time τ



Markov models

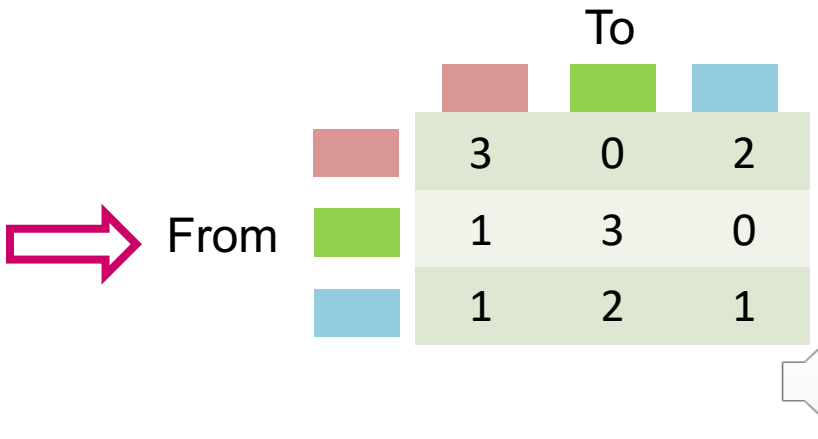
Total sampled time (e.g. 100 μs)

Define the *discrete* state of the system in *discrete time* (e.g. via reaction coordinates)

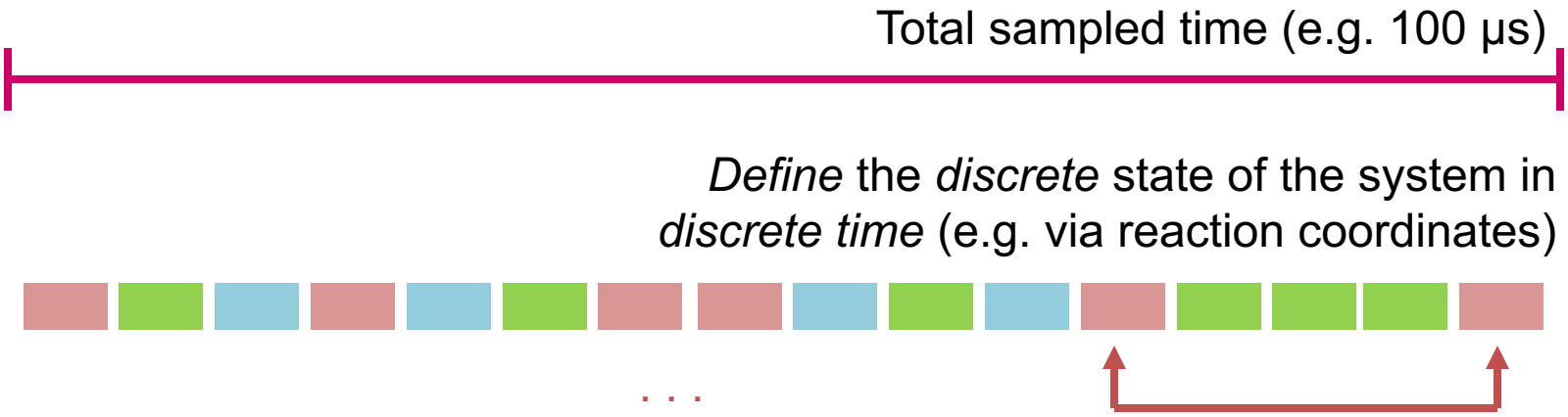


Lag time τ (here: 4 time units)

Compute the transition probability matrix sliding a window of lag time τ



Markov models



Transition counts

From \ To			
	3	0	2
	1	3	0
	1	2	1

Normalize by rows

Transition probabilities

From \ To				Σ_j
	3/5	0	2/5	1
	¼	¾	0	1
	¼	2/4	¼	1

P_{ij}

Repeat until the end of the trajectory.

Note the Markovian assumption: transition probabilities **do not** depend on history (neither, for us, on time).

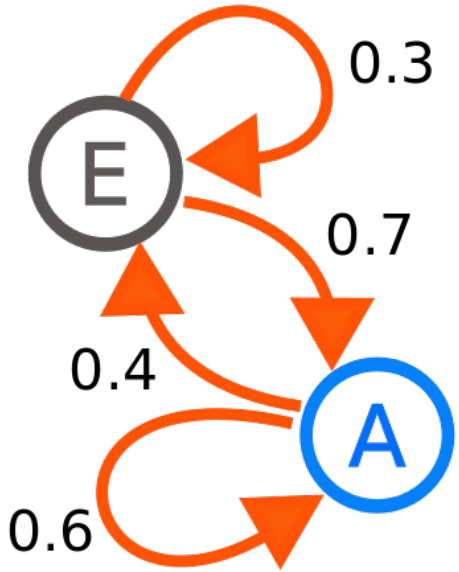
(Note how "Early" and "late" events are squashed in the same matrix.)

Discrete-time Markov Chains

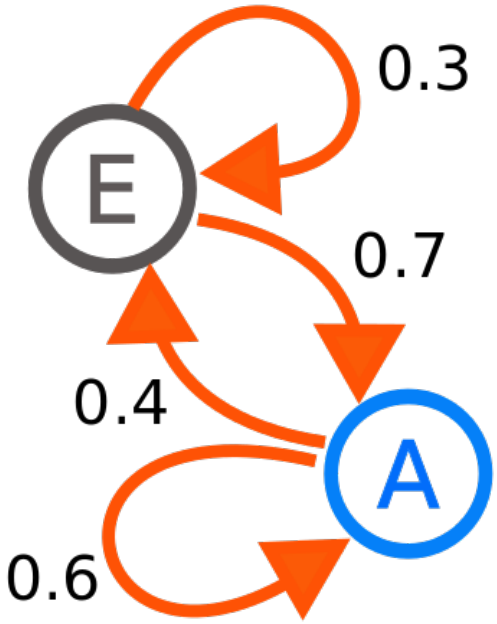
Andrei Markov
1856-1922

Discrete Time Markov Chains

- A **random** process.
- The system's state is a **discrete** variable.
- It undergoes transitions between states at uniformly-spaced (**discrete**) time points.
- Transition probabilities do not depend on the previous history of states (**memorylessness**).



Transition probability matrix



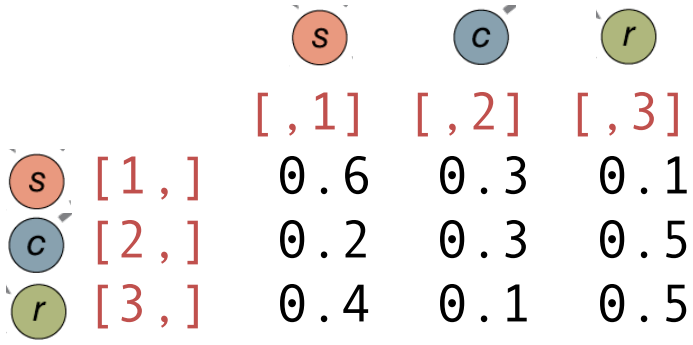
P_{ij} = *Probability to change state from i to j at each time point*

$$P_{ij} = P(X_t = j | X_{t-1} = i)$$

i \ j	A	E
A	0.6	0.4
E	0.7	0.3

Let's make one with 3 states…

Q: Can you match state names with matrix indices?

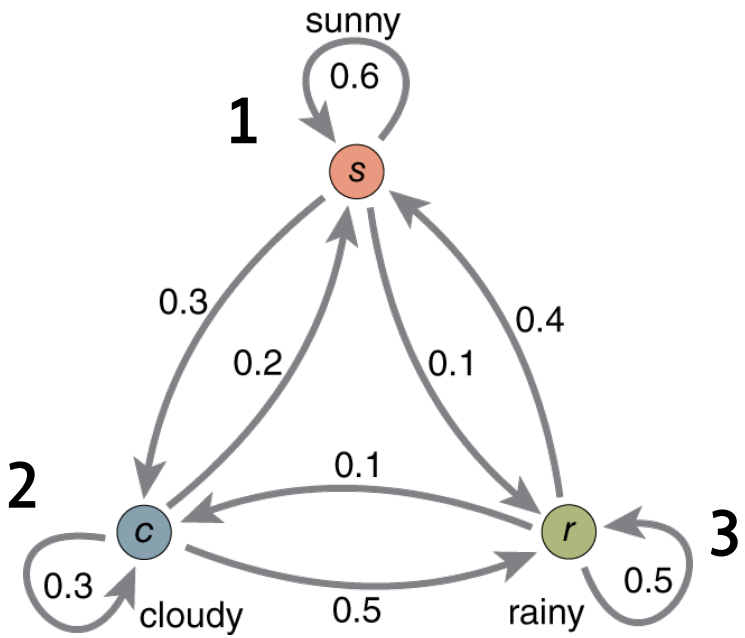


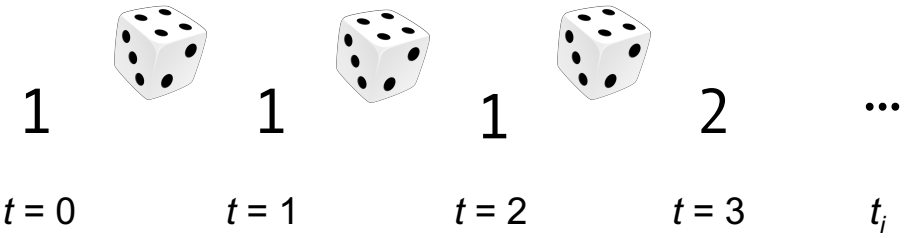
		s	c	r
		[,1]	[,2]	[,3]
s	[1,]	0.6	0.3	0.1
c	[2,]	0.2	0.3	0.5
r	[3,]	0.4	0.1	0.5

Let's generate samples…

Start e.g. from state 1
(we'll use numbers from now on).

Play the game and follow the Markov chain for many (discrete) time steps.

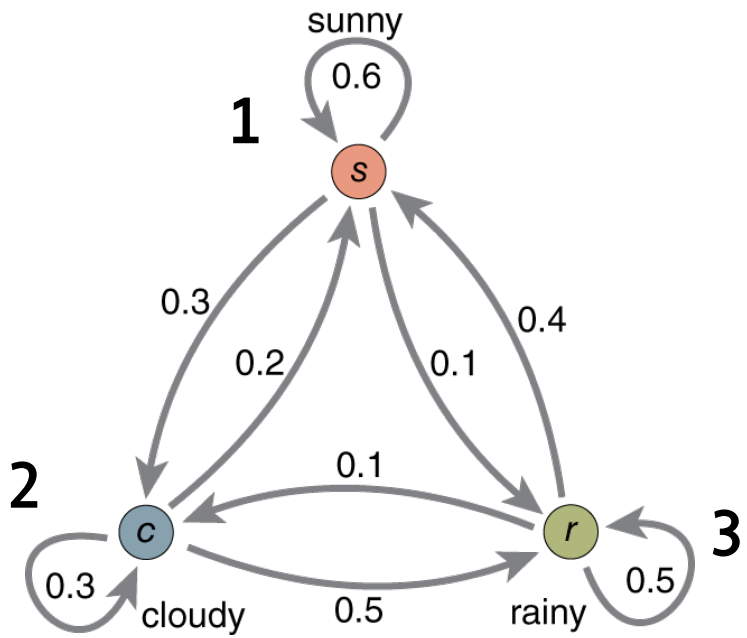




1 1 1 2 2 3 3 1 1 1 2 2 2 2 1 1 1 1 1 1 3 1 1 1 1 2 3 1 1 3...

https://github.com/giorginolab/Markov-Tutorial-BCN-2021

Equilibrium probabilities



Question: if we play the game *a very long* time, what fraction of time would we spend in each state?

- That is the *asymptotic* (equilibrium) distribution.
- I.e., the probability of finding the system in a given state.

1 1 1 2 2 3 3 1 1 1 2 2 2 2 1 1 1 1 1 1 3 1 1 1 1 2 3 1 1 3...

(1000 times)

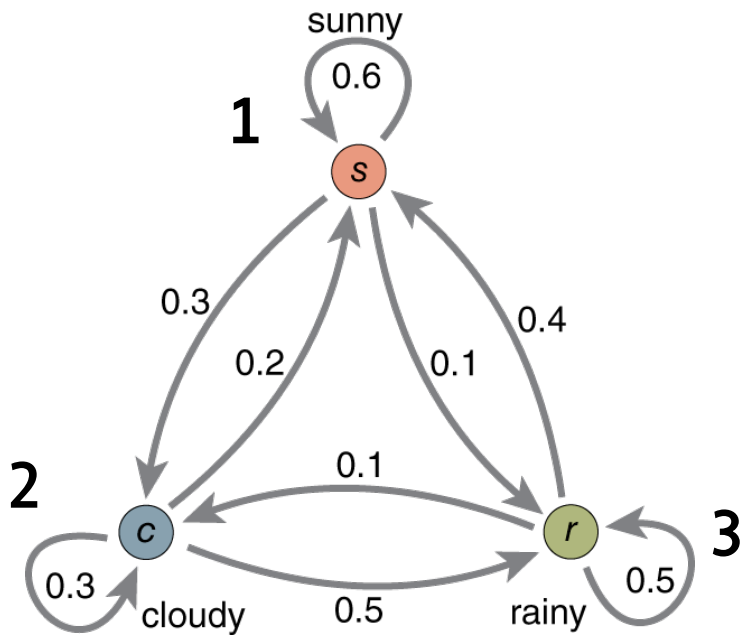
```
1    2    3
431  220  349
```

$p_\infty(1) = 0.431$
$p_\infty(2) = 0.220$
$p_\infty(3) = 0.349$

$\Sigma_i\, p_\infty(i) = 1$

https://github.com/giorginolab/Markov-Tutorial-BCN-2021

Equilibrium probabilities



Question: if we play the game *a very long* time, how much time would we spend in each state?

- That is the *asymptotic* (equilibrium) distribution.
- I.e., the probability of finding the system in a given state.
- I.e., the free energy (ΔG) of that state!

1 1 1 2 2 3 3 1 1 1 2 2 2 2 1 1 1 1 1 1 3 1 1 1 1 2 3 1 1 3...

(1000 times)

```
1    2    3
431 220 349
```

$p_\infty(1) = 0.431$
$p_\infty(2) = 0.220$
$p_\infty(3) = 0.349$

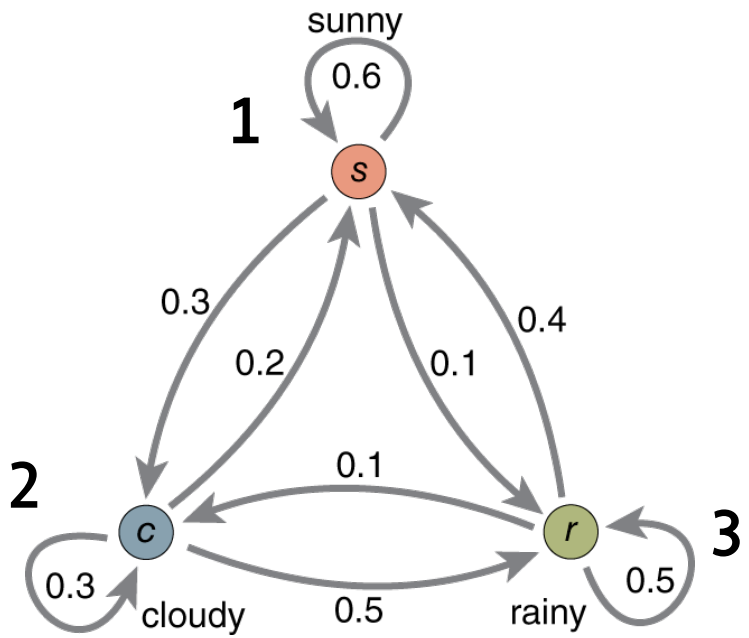
$\Sigma_i\, p_\infty(i) = 1$

$G_1 = -k_B T \log(p_1) \simeq 0.50$ kcal/mol
$G_2 = -k_B T \log(p_2) \simeq 0.91$ kcal/mol
$G_3 = -k_B T \log(p_3) \simeq 0.63$ kcal/mol

$\Delta G_{11} \doteq 0$ kcal/mol
$\Delta G_{21} \simeq 0.37$ kcal/mol
$\Delta G_{31} \simeq 0.18$ kcal/mol

https://github.com/giorginolab/Markov-Tutorial-BCN-2021

Equilibrium probabilities



Question: if we play the game *a very long* time, how much time would we spend in each state?

- That is the *asymptotic* (equilibrium) distribution.
- I.e., the probability of finding the system in a given state.
- I.e., the free energy (ΔG) of that state!

~~1 1 1 2 2 3 3 1 1 1 2 2 2 2 1 1 1 1 1 1 3 1 1 1 1 2 3 1 1 3...~~

~~(1000 times)~~

Sample and count

Compute eigenvectors of matrix *P*

```
     [,1] [,2] [,3]
[1,]  0.6  0.3  0.1
[2,]  0.2  0.3  0.5
[3,]  0.4  0.1  0.5
```

→ Compute eigenvectors: same result in one shot

https://github.com/giorginolab/Markov-Tutorial-BCN-2021

Intermezzo

Google's PageRank

SIAM Review
Vol. 48, No. 3, pp. 569–581

The $25,000,000,000 Eigenvector: The Linear Algebra behind Google*

Kurt Bryan†
Tanya Leise‡

Abstract. Google's success derives in large part from its PageRank algorithm, which ranks the importance of web pages according to an eigenvector of a weighted link matrix. Analysis of the PageRank formula provides a wonderful applied topic for a linear algebra course. Instructors may assign this article as a project to more advanced students or spend one or two lectures presenting the material with assigned homework from the exercises. This material also complements the discussion of Markov chains in matrix algebra. Maple and *Mathematica* files supporting this material can be found at www.rose-hulman.edu/~bryan.

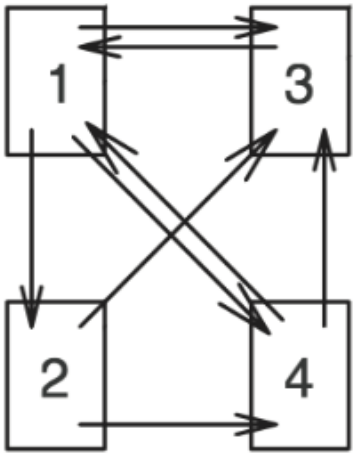
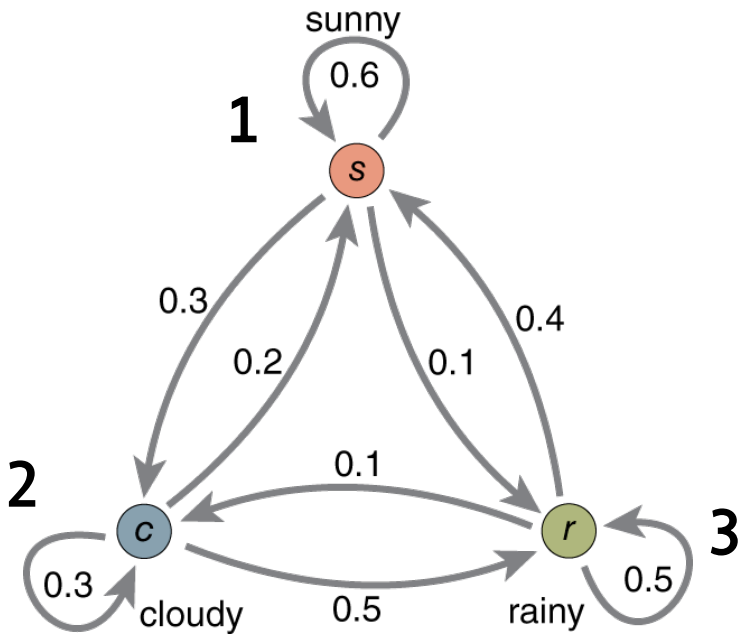


Fig. 1 *An example of a web with only four pages. An arrow from page* A *to page* B *indicates a link from page* A *to page* B.

Kinetics

Question: if we start from 1, how many steps *on average* do I wait to reach 2?

- This is the 1 → 2 *mean first passage time*
- I.e., the (inverse) transition rate ($1/k_{on}$) between those states!



1 1 1 2 2 3 3 1 1 1 2 2 2 2 1 1 1 1 1 1 3 1 1 1 1 2 3 1 1 3…

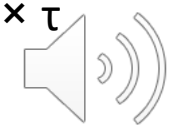
3 2 1 3 2 1 11

And so on…

Answer: average all the numbers in red.

(Actually done mathematically on P_{ij})

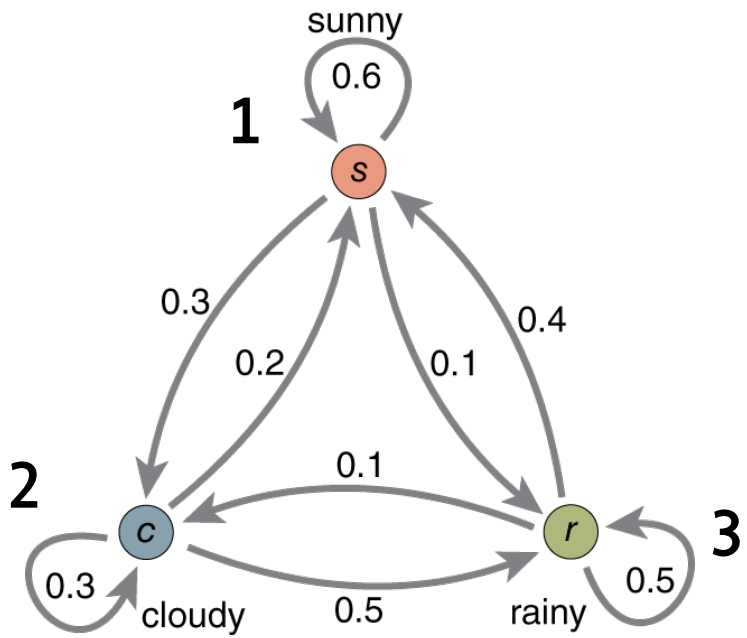
- Obviously one can ask the same question about any $i \rightarrow j$
- Remember that time is steps × τ



Committor

Question: if we start from 1, what is the probability to pass through 2 w.r.t. 3 first? (regardless of the path)

- This is the 1 → {2,3} *committor probability*
- States which are equally probable to fall back to the "reactants" or advance to "product" state form the *transition state*.



1 1 1 2 2 3 3 1 1 1 2 2 2 2 1 1 1 1 1 1 3 1 1 1 1 2 3 1 1 3…

2
2
2
2
2
2
3
3

And so on…

Answer: compute the fraction of end-in-2 vs end-in-3.

(Actually done mathematically on P_{ij})

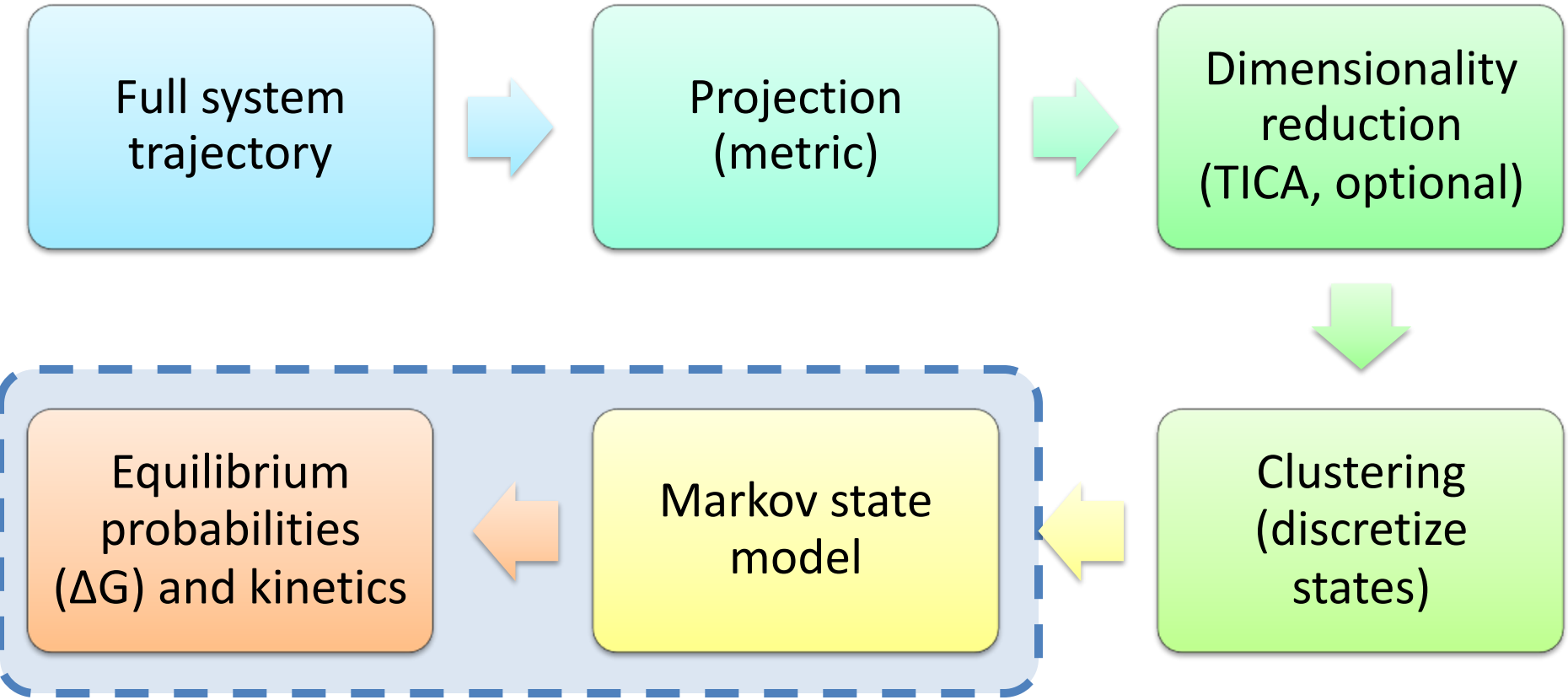
Markovianity

- The state transition probabilities only depend on the current state. Examples:
  - Today's weather, not yesterday's
  - Where the ligand is, not how did it got there
- The property may be false at short timescales but true at longer ones: study varying τ
- It does depend on the chosen states: study varying state discretization

Back to molecules:

>> 1 dimension

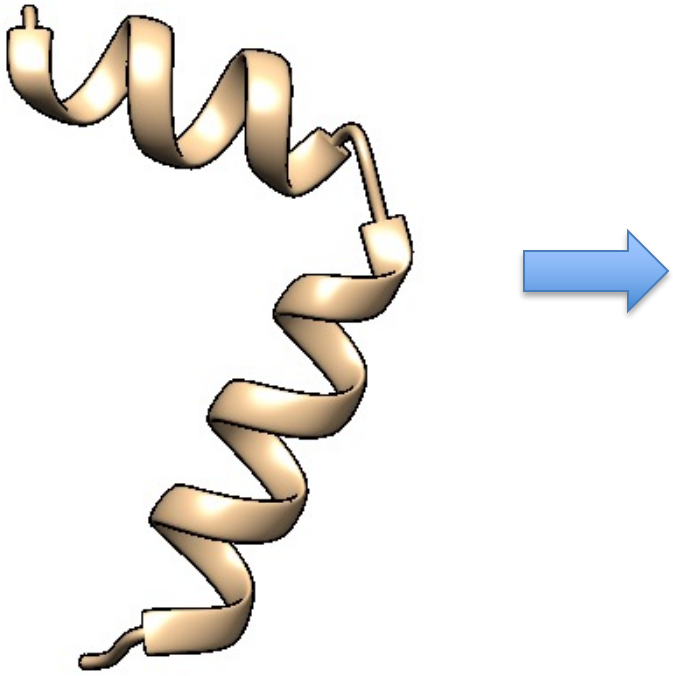
MSM-based analysis overview



MSM proper – this is the part we have seen so far

Metric projection

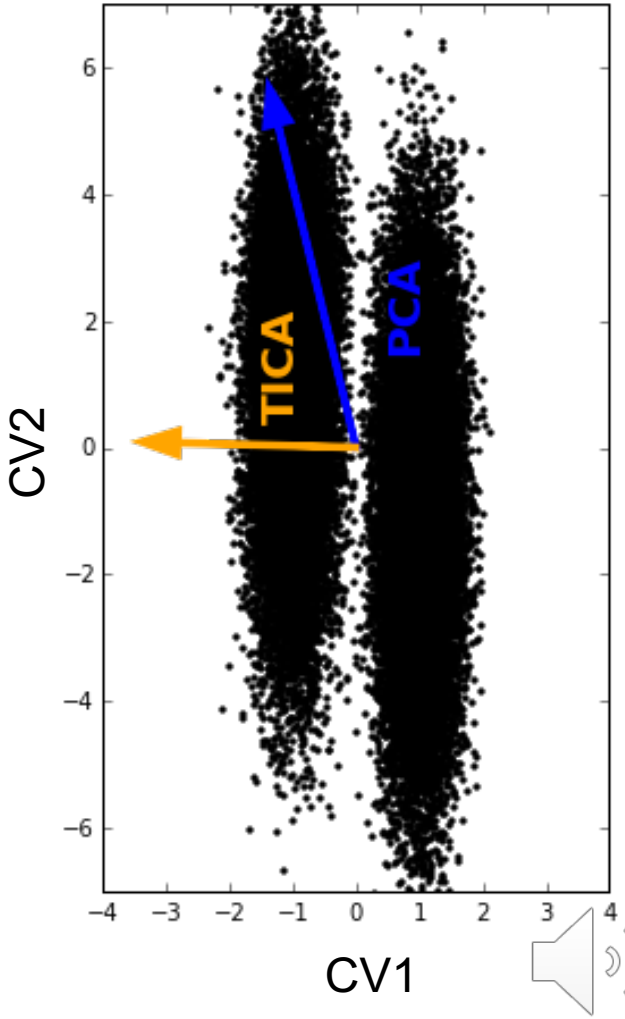
The first step is to project the system state in a lower-dimensional space ("metric"). Many choices are available, e.g.



- Manually chosen distances
- Atom coordinates
- *N* phi/psi Ramachandran angles
- Distance matrix
- Contact matrix
- …

Time-lagged independent component analysis

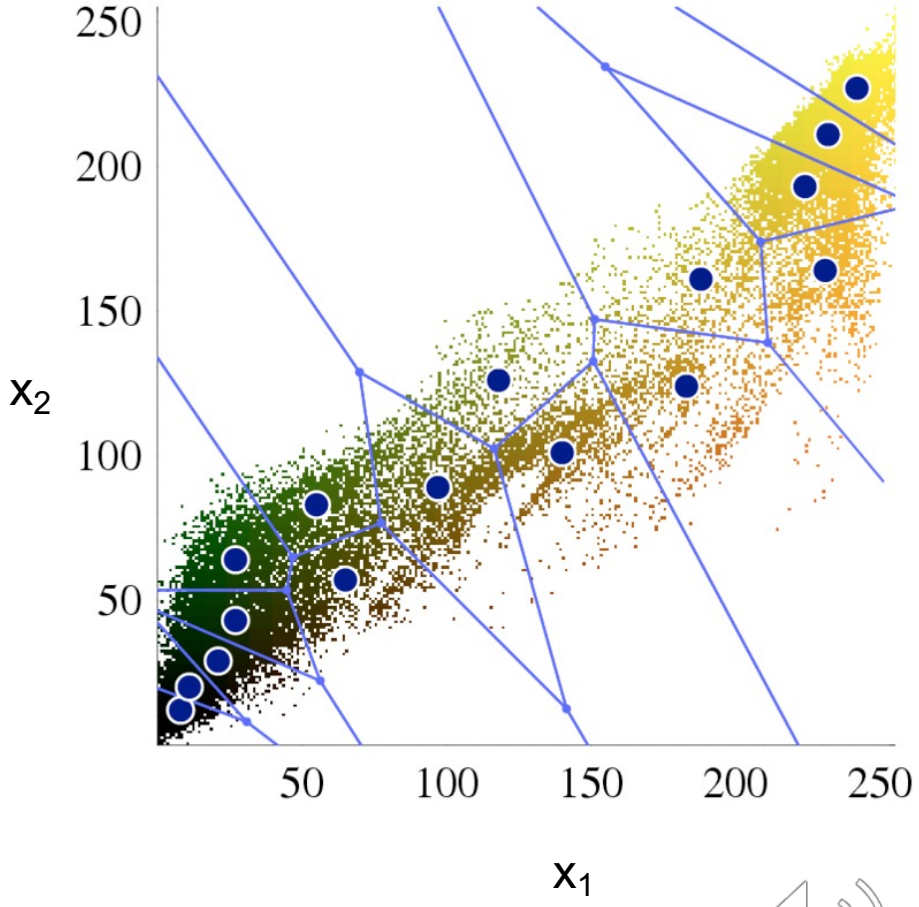
- The feature space dimensionality may still be too high
- Do a further low-dimensional projection on the "slow" degrees of freedom: TICA
- It is based on lagged autocorrelation
- Contrast with PCA, which fits the "most elongated" ellipsoid, ignoring time



J. Chem. Phys. 139, 015102 (2013).

Clustering (1st level)

- Now move to a discrete state space
- Reduction from the low-dimensional space is done by *clustering*
  - Usually called "microstates"
- Several algorithms are implemented in MSM packages (e.g. grid; *k*-means; etc. We won't discuss them)



https://en.wikipedia.org/wiki/K-means_clustering

Examples from the literature

Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations

Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis[1]

Computational Biochemistry and Biophysics Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park, C/Doctor Aiguader 88, 08003 Barcelona, Spain

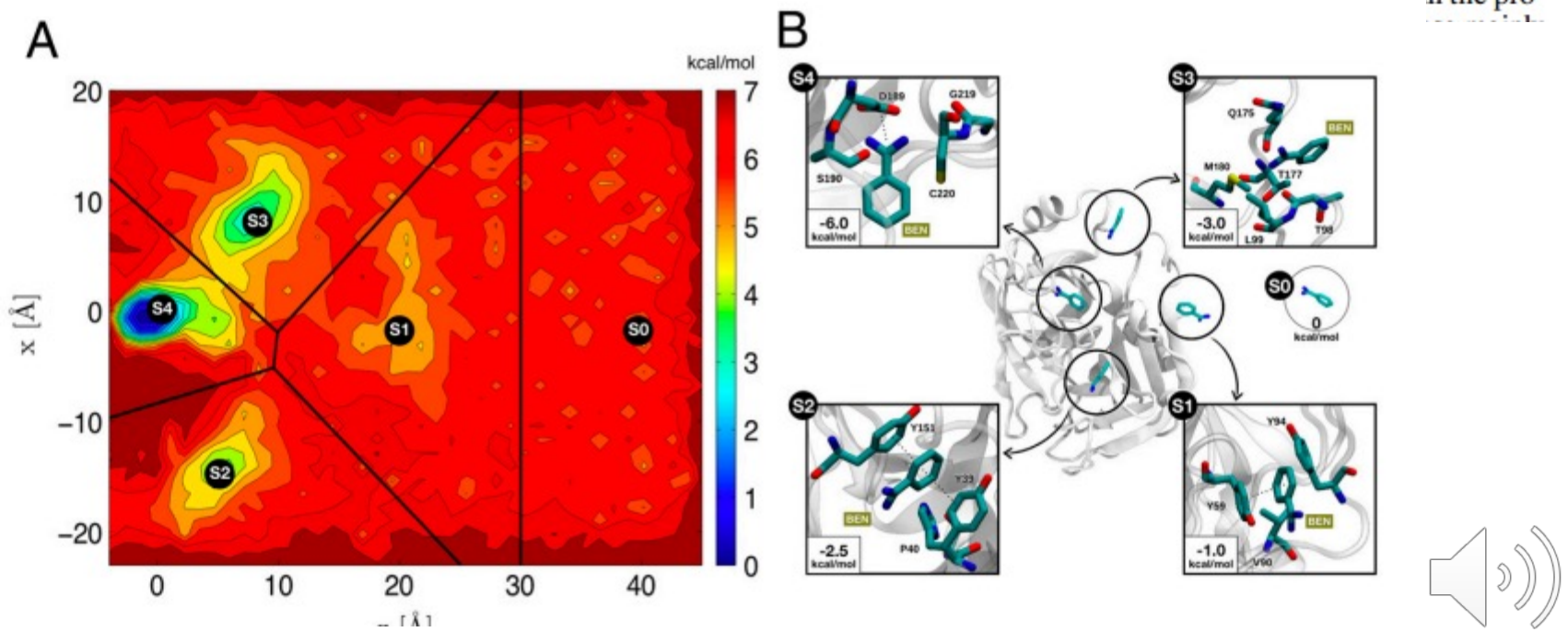
Edited by Arieh Warshel, University of Southern California, Los Angeles, CA, and approved May 11, 2011 (received for review March 4, 2011)

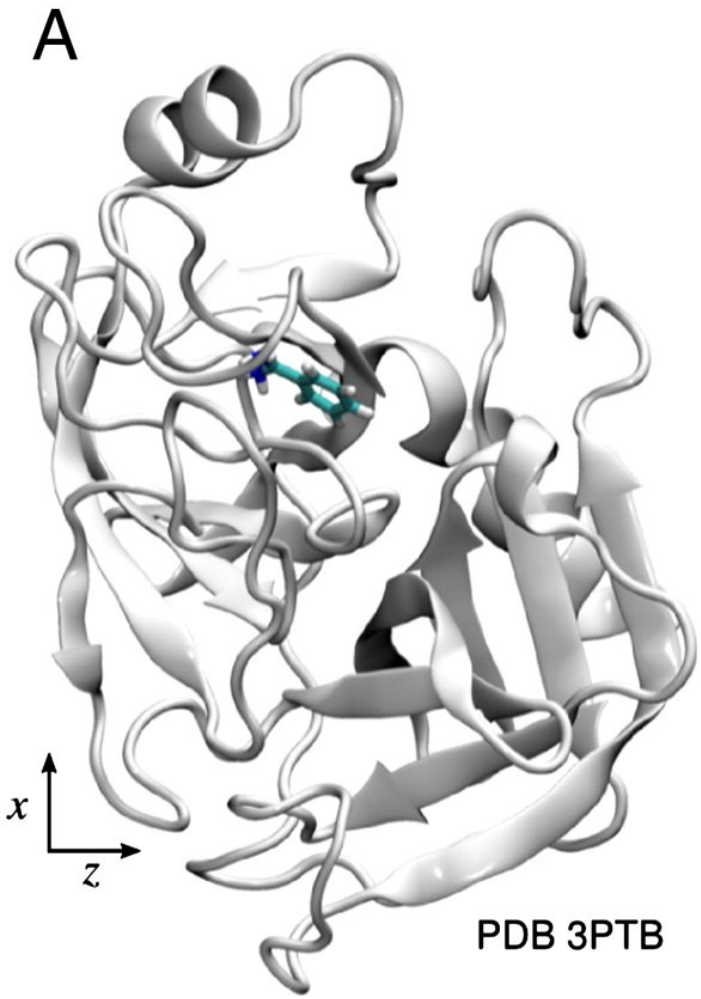
The understanding of protein–ligand binding is of critical importance for biomedical research, yet the process itself has been very difficult to study because of its intrinsically dynamic character. Here, we have been able to quantitatively reconstruct the complete binding process of the enzyme-inhibitor complex trypsin-benzamidine by performing 495 molecular dynamics simulations of fr

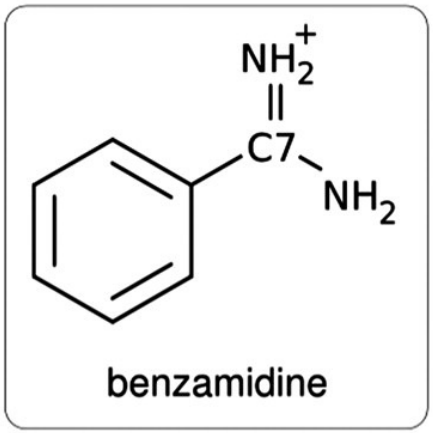
reproduce with atomic resolution the crystallographic mode of binding, but we also provide the kinetically and energetically meaningful transition states of the process.

Free ligand binding has been used in the past to describe computational experiments in which, typically, a ligand is placed at a certain distance from the target protein and first by diffusion and

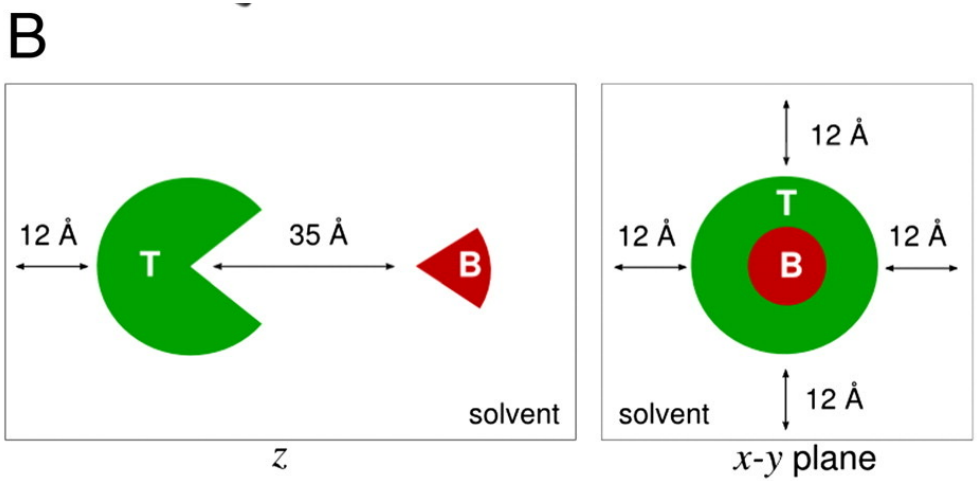
n the pro-



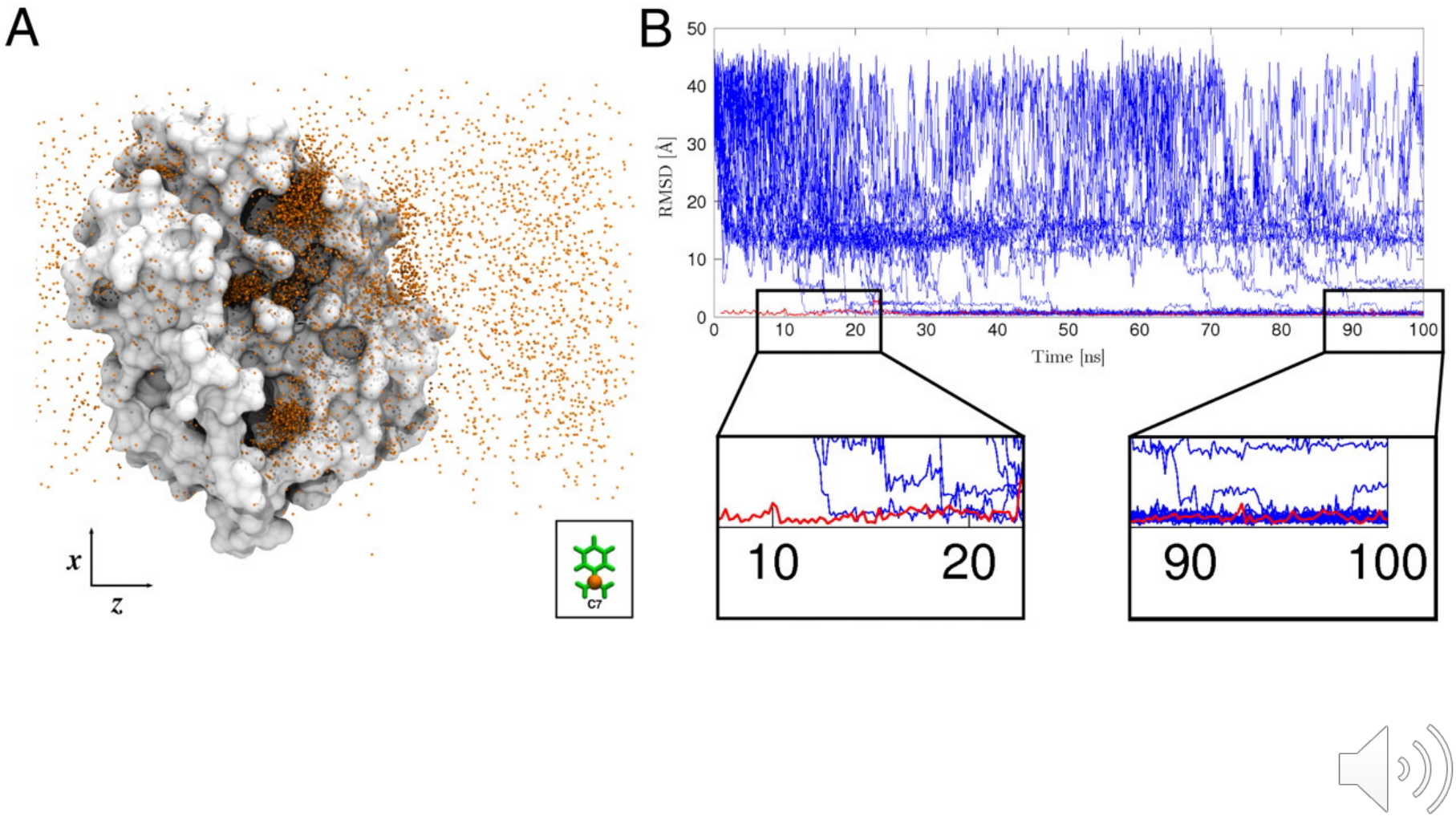


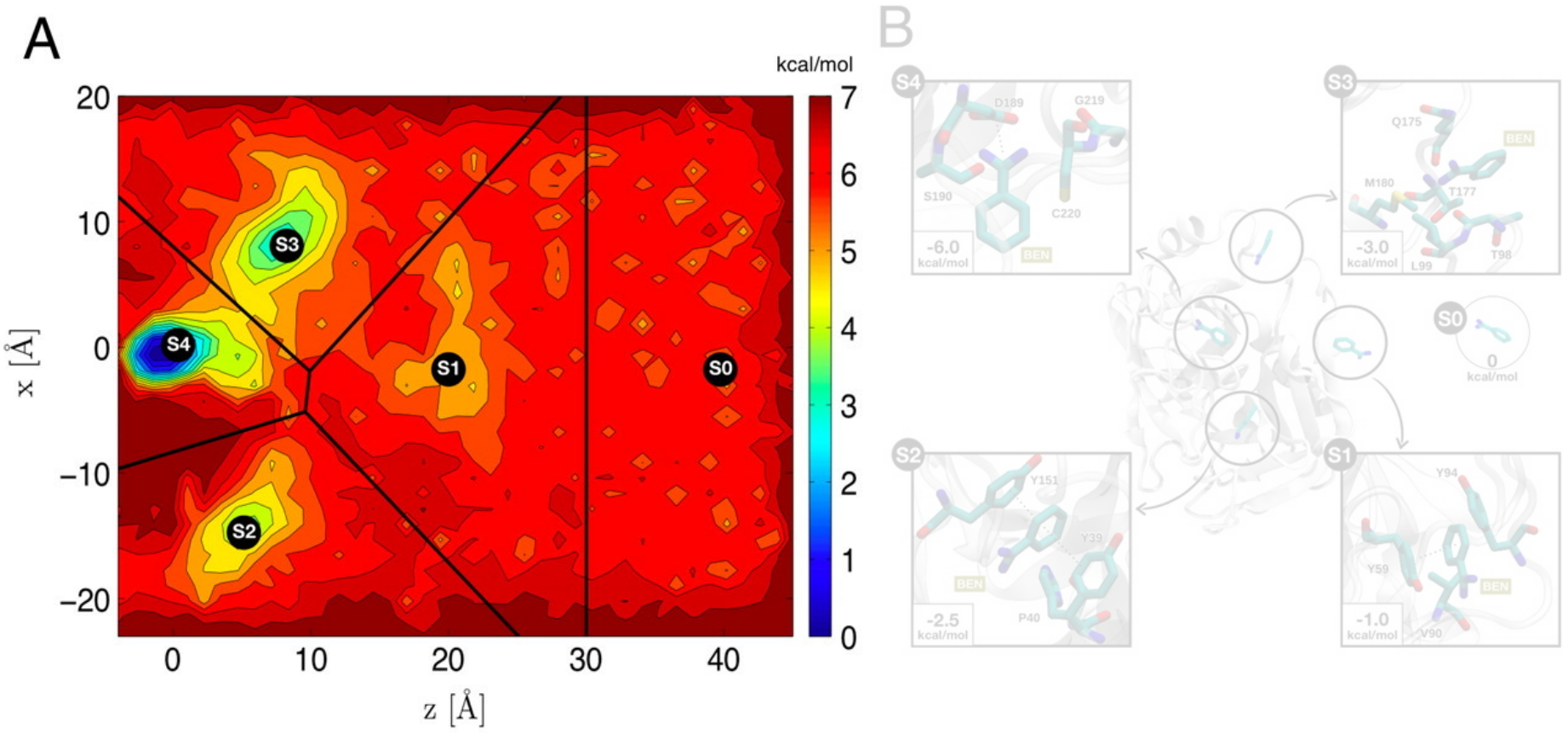


- Rigid ligand: no internal degrees of freedom
- Small ligand: no orientation d.o.f.
- States defined as small cubes according just to the position in 3D space



Ignasi Buch et al. PNAS 2011;108:10184-10189


A
x
z
C7
B
RMSD [Å]
Time [ns]
0
10
20
30
40
50
60
70
80
90
100
10
20
90
100

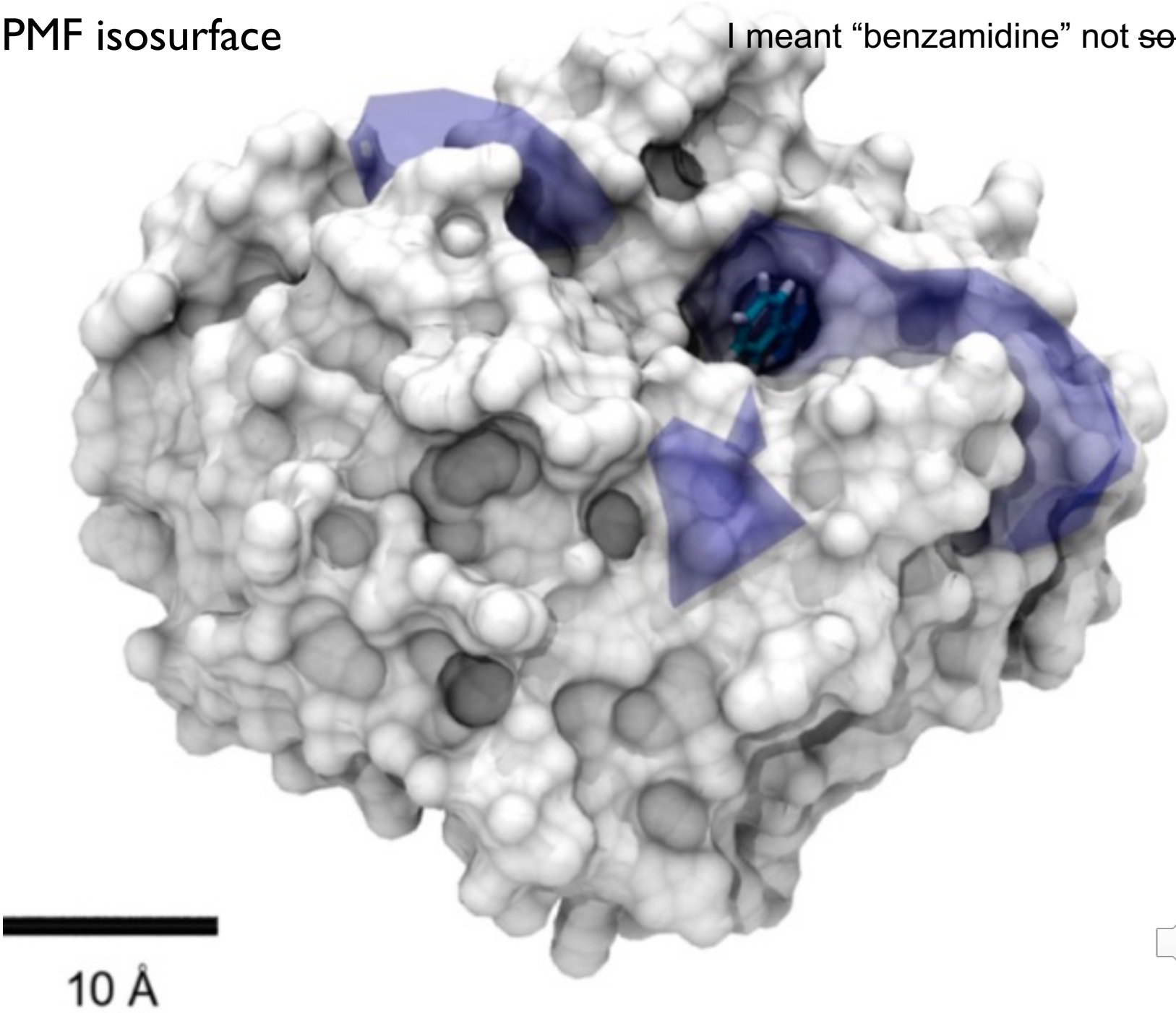


Identification of metastable states. (A) PMF in the xz plane. Five different metastable states can be identified from the different free-energy minima (S0 to S4). The relative free energy between the unbound state S0 and the bound state S4 is -6 kcal/mol. The most probable transition to the bound state S4 may be from S3 from the fact that the barrier between the two states is just 1.5 kcal/mol. (B) Structural characterization of metastable states. In states S1 and S2, benzamidine is stabilized by π-π stacking interactions with Y151 and Y39 side chains. In S3, a hydrogen bond may be formed between NH2 groups of benzamidine (only heavy atoms shown for clarity) and Q175 side chain, or by a cation-π interaction between the Q175 side chain again, and benzamidine's benzene ring.

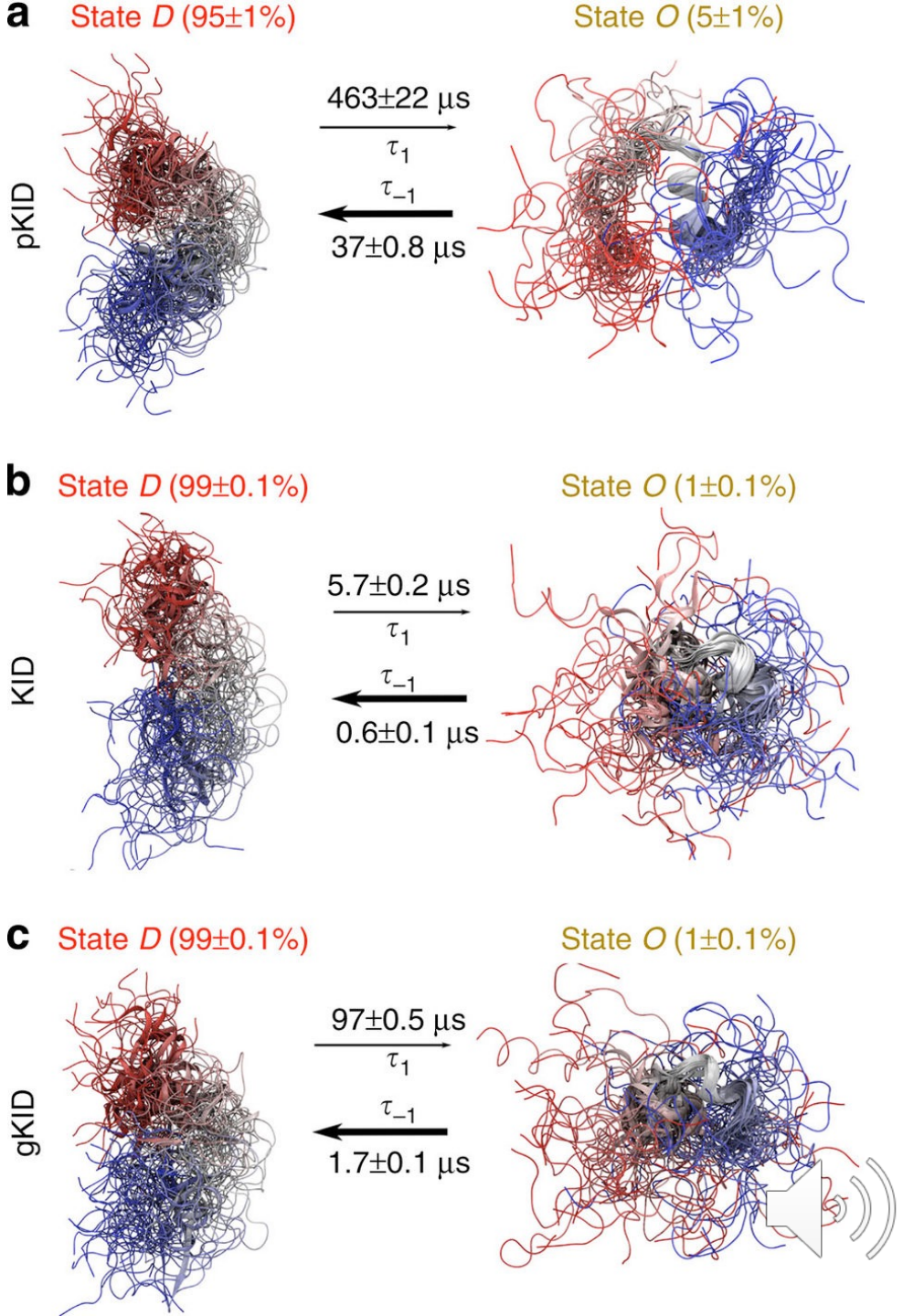
Ignasi Buch et al. PNAS 2011;108:10184-10189

3D PMF isosurface

I meant “benzamidine” not ~~solvent~~



Kinetic modulation of a disordered protein domain by phosphorylation



N. Stanley, S. Esteban and G. De Fabritiis, Nat. Commun. 5, 5272 (2014)

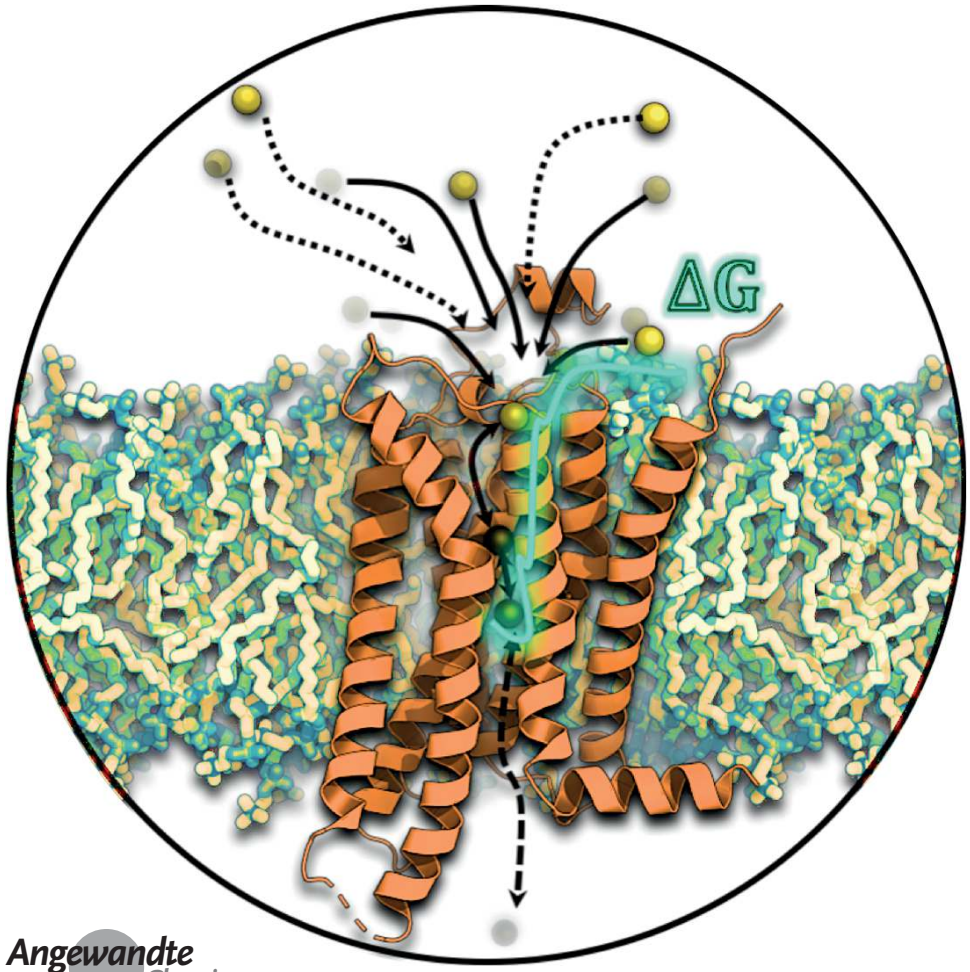
doi:10.1038/ncomms6272

Allosteric Modulators

International Edition: DOI: 10.1002/anie.201708889
German Edition: DOI: 10.1002/ange.201708889

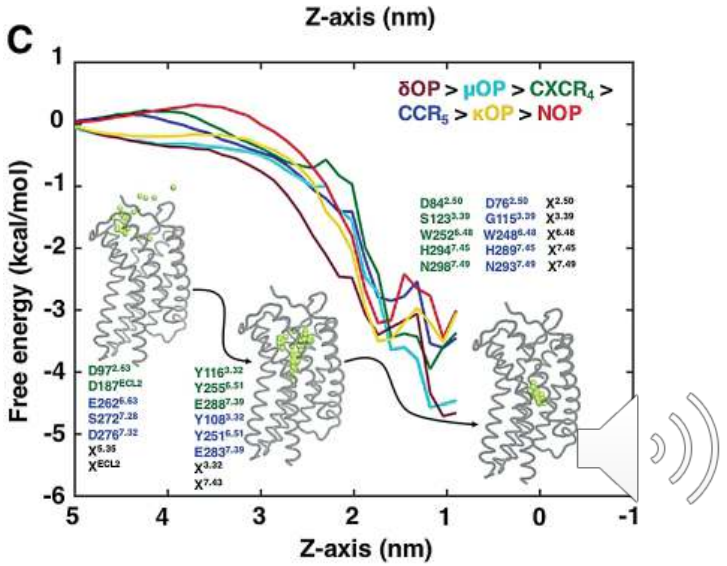
Universality of the Sodium Ion Binding Mechanism in Class A G-Protein-Coupled Receptors

*Balaji Selvam, Zahra Shamsi, and Diwakar Shukla**



plays a key role in identifying drug candidates.[7] Several studies have shown that GPCRs are allosterically modulated by endogenous Na^+ ions.[8–11] Selent and co-workers explored the binding mechanism of Na^+ to the D_2 receptor by molecular dynamics simulations.[12] Na^+ binds in the middle of the transmembrane (TM) region to a conserved $Asp^{2.50}$ residue (Ballesteros–Weinstein numbering),[13] as elucidated by high-resolution crystal structures of the GPCRs.[14–16] More recent MD studies on μ-, κ-, and δ-opioid (OP) receptors also support Na^+ ion binding to $Asp^{2.50}$.[17,18] Na^+ binding and a change in the protonation states of titratable residues through the water network have a significant effect on the stabilization of the conformational states of GPCRs.[19–21]

elusive. Herein, we performed hundreds-of-microsecond long simulations of 18 GPCRs to elucidate their Na^+ binding mechanism (see the Supporting Information, Tables S1 and S2). We constructed Markov state models (MSMs) to estimate the free energy profiles for Na^+ binding to each GPCR. Analysis of the Na^+ binding kinetics revealed key residues that act as major barriers for Na^+ entry to the intracellular site. We also predicted the average mean first passage time (MFPT) for Na^+ binding and unbinding events by transition path theory (TPT; Table S3).



Conclusions and Resources

Warning

- Still very active field
- Suggested further steps: worked out real-world examples distributed with software packages*
- For serious work, many more details are in…
  - the theory of Markov state models
  - the discretization, projection, and estimation of models from trajectories

* see the last slides

Conclusions

- MSM methods may be an attractive formalism for medium-sized problems
- Make efficient use of *unbiased* sampling
- Still require *huge* (but achieavable) amounts of sampling/simulation for biologically interesting systems
- Strong mathematical foundation, with good software available

Software + Tutorials

- All are Python-based
  - They include clear walkthroughs (highly recommended) with datasets
- PyEMMA – www.emma-project.org
- HTMD – www.htmd.org
  - Also analysis + system build + adaptive …
  - Can aggregate large-scale datasets
- MSMBuilder - msmbuilder.org

Literature

- Buch et al., 10.1073/pnas.1103547108
  - Rigid ligand + protein association, simplest case
- Swinney, PMID:19152211
  - Importance of kinetics in drug design
- Chodera and Noe, 10.1016/j.sbi.2014.04.002
  - Excellent overview (1)
- Voelz et al., 10.1021/ja9090353
  - Full reconstruction of a millisecond folding
- Selvam et al., 10.1002/anie.201708889
  - Reconstruction of Na++ binding to GPCRs
- Pande et al., 10.1016/j.ymeth.2010.06.002
  - Excellent overview (2)
- Paul et al., 10.1038/s41467-017-01163-6
  - Peptide-peptide association
- Doerr et al., 10.1021/ct400919u
  - Adaptive sampling
- Bryan et al., "The $25 B eigenvector"
  - The original PageRank algorithm; e.g. jdc.math.uwo.ca/M1600b/l/pagerank-1600.pdf

End