Markov-state modeling of biomolecular systems

GitHub
@tonigi

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

Structural Bioinformatics @UniMi
Jan 15, 2019

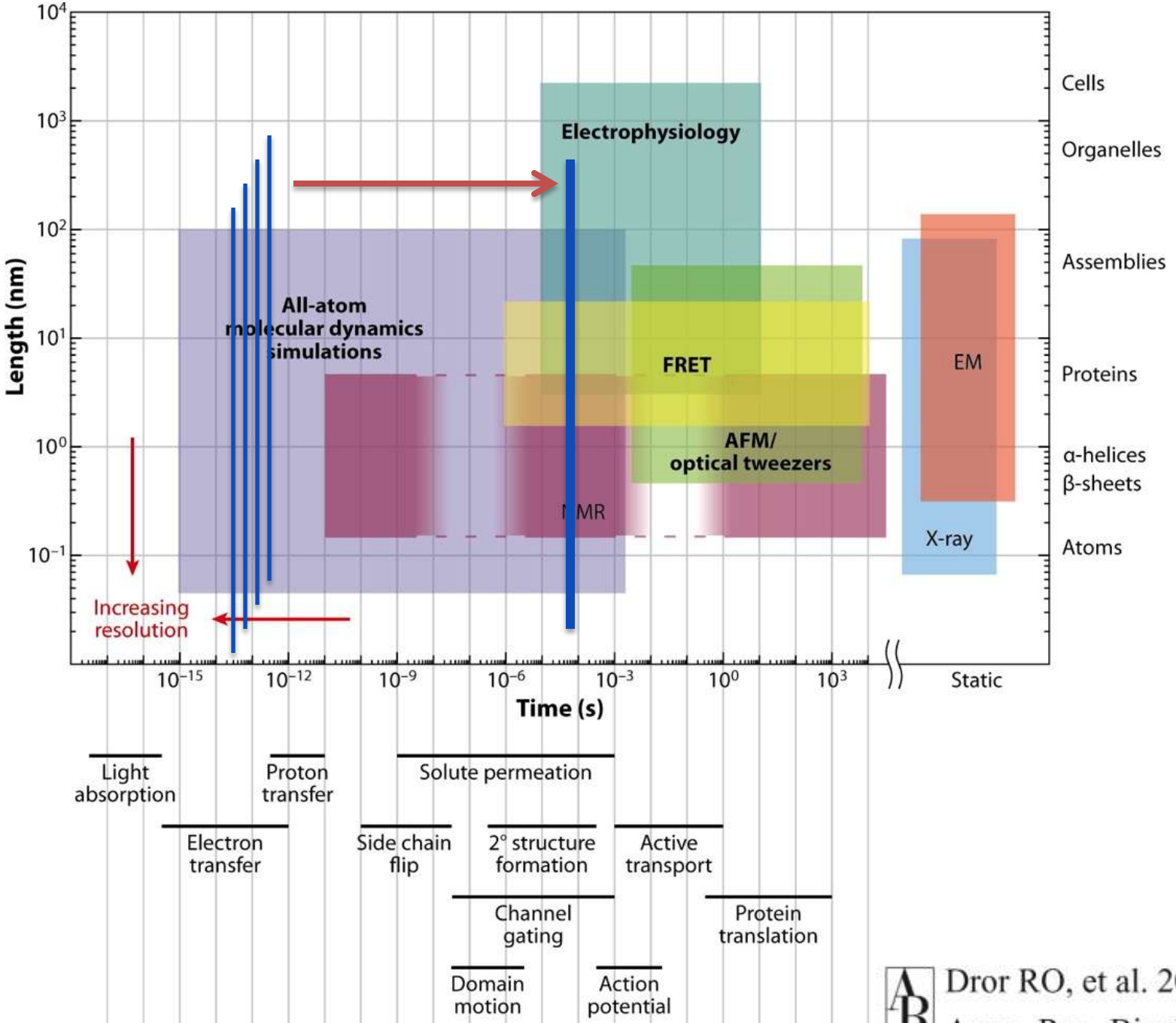
Introduction

- 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?



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

Warning

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

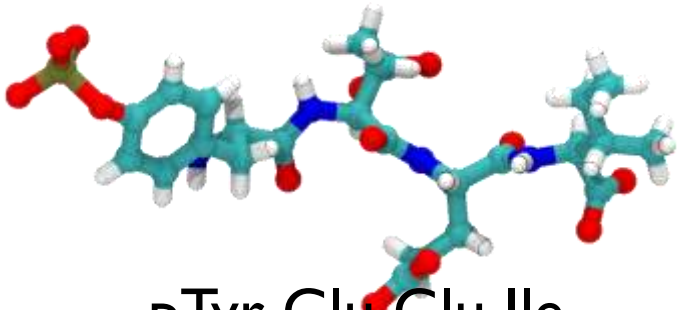
* see the last slides

Binding as an irreversible two-state model

Unbiased simulations:
SH2 – pYEEI binding example

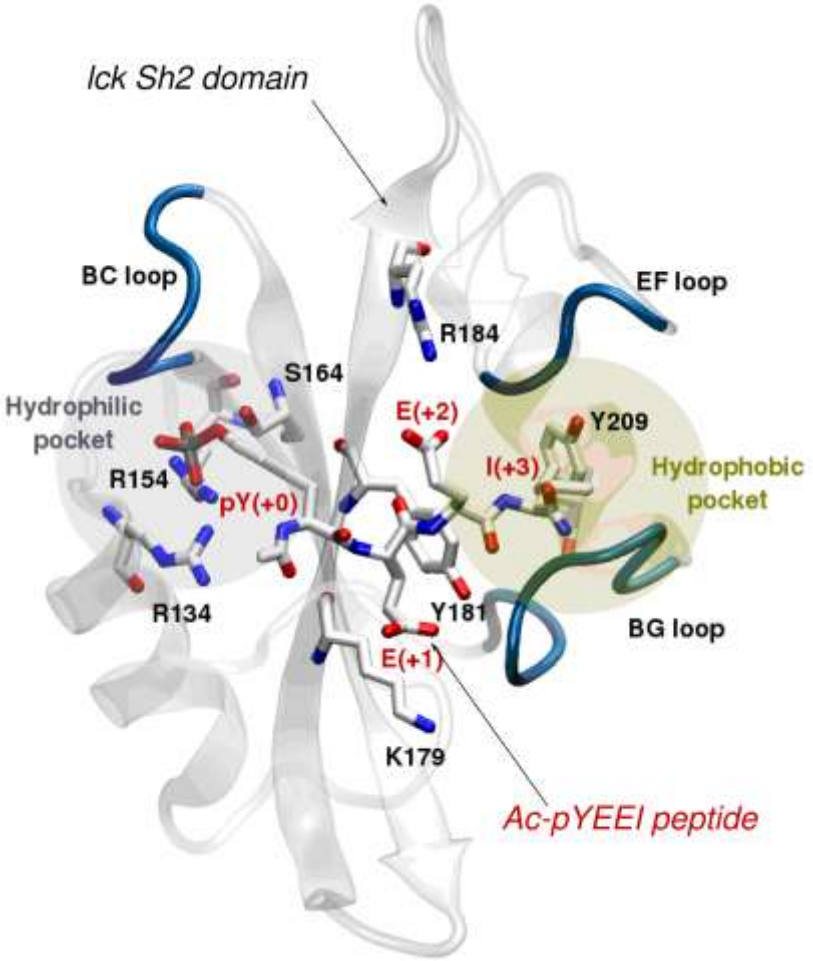
Giorgino et al., JCTC 2012

SH2 Domains

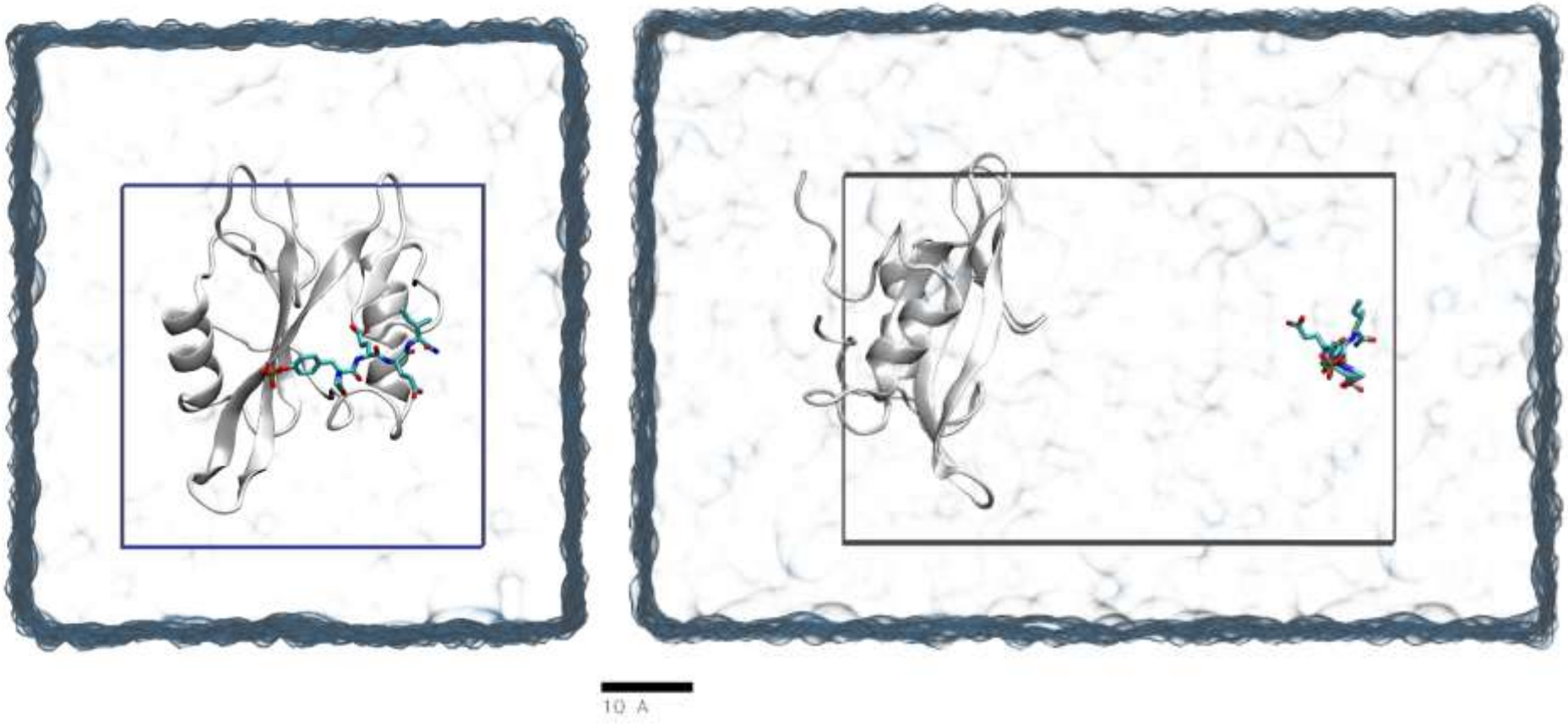


pTyr Glu Glu Ile

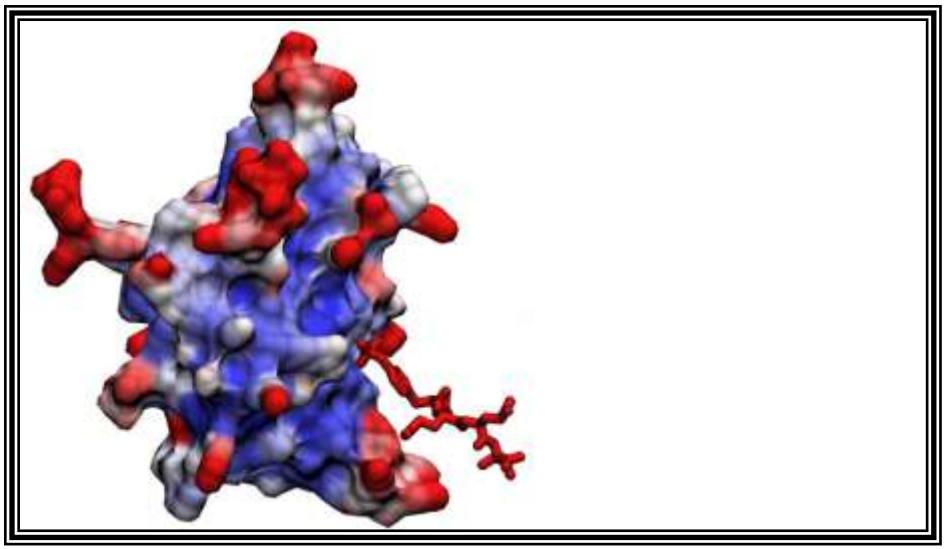
- Small domain (100 AA), well-conserved
- Recognize short pY-* sequences with high specificity
- Prominent role in signaling
- Found in ~110 human proteins
- "Socket with two holes"



In silico set-up



Teaser (1:36)



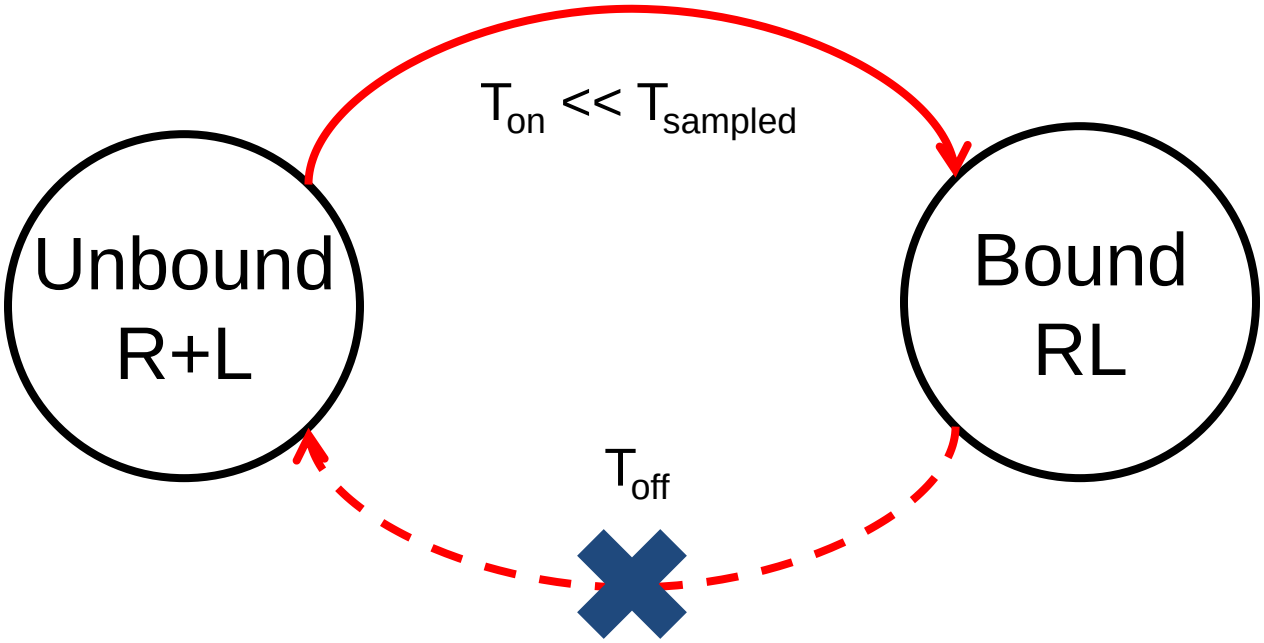
$$R + L \rightarrow RL$$

http://goo.gl/ZLI42
also https://youtu.be/Xhof-pf_Eo8

- *Video: unbiased binding simulation of pYEEI to Lck SH2 domain.*
- *Protein surface and ligand's atoms color-coded by local flexibility, from blue to red (lowest to highest RMSF).*
- *As the peptide binds into its native conformation (RMSD < 2 Å from crystal structure), it loses flexibility; the same happens to the protein.*
- ***By construction, we are not at equilibrium.***

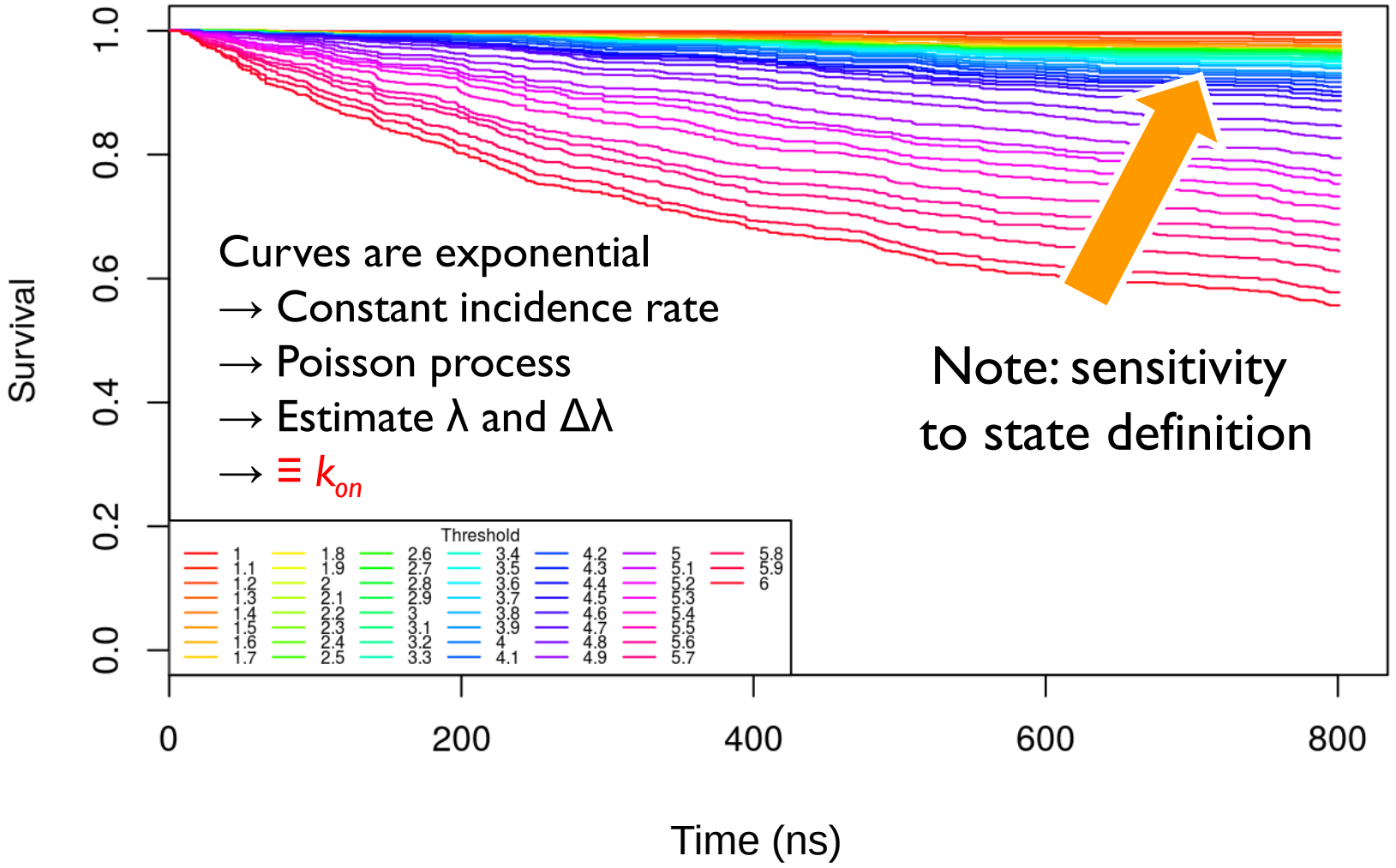
Binding as an *irreversible* two-state model

On the sampled timescales…



$$T_{event} << T_{on} << T_{sampled} << T_{off}$$

Non-bound-fraction curves



Motivation for a more general solution

This 2-state method is not sufficient:

1. One needs to define the *bound state*
   – which is not available if we lack
   a crystallographic or NMR structure.
2. We need to observe a sizeable number of *full* binding events.
   – Association processes are slow.
   They may be (far) outside our sampling power.
3. We get no information about *off* rates.
4. We get no information about *metastable states*.

Going multi-state: Discrete-time Markov Chains

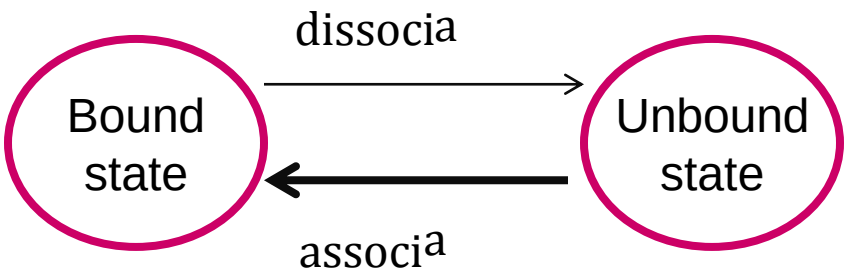
Why do we care?

Because equilibrium probabilities *are* the free energy of binding (ΔG).

ΔG is an important determinant of drug potency.

$$K_D = p_U/p_B \, (\propto [P][L]/[PL])$$
$$= \exp(-\Delta G/RT)$$

$$\Delta G = -RT \ln K_D$$

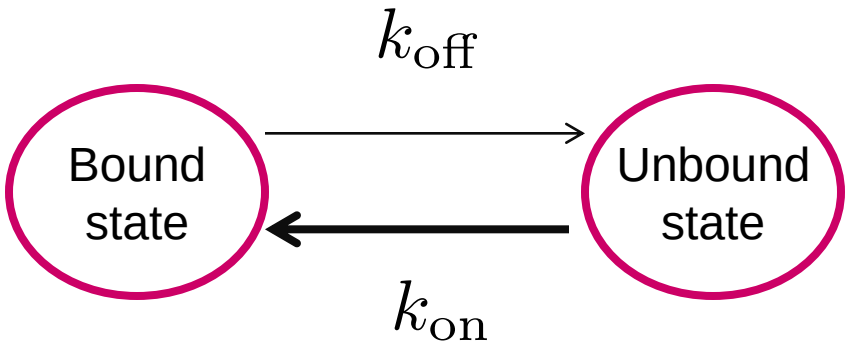


Kinetics

But the real breakthrough is that we can use the *dynamic* information, computing the *kinetics* as well. This is why we are doing *unbiased dynamics* in the first place!

Binding kinetics is important for rational drug design *as well* (e.g. related to the time spent in complex with receptors).

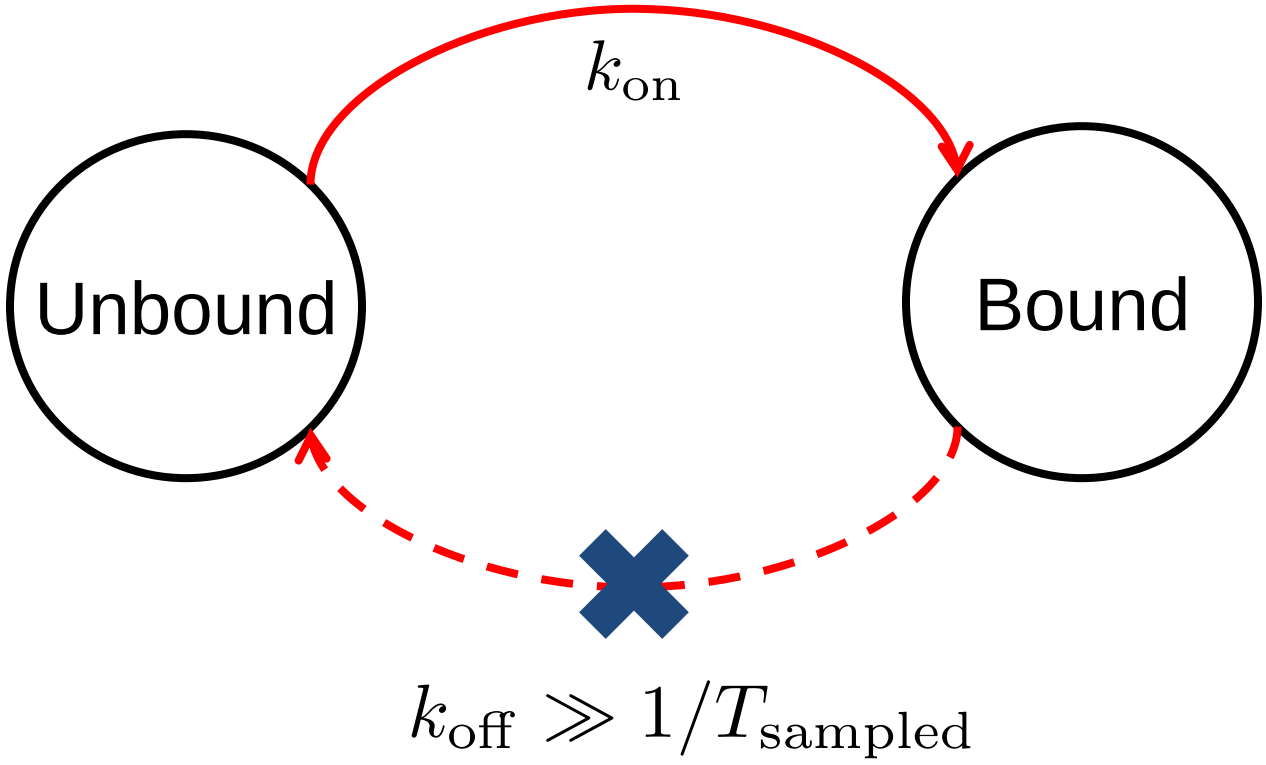
$$K_D = k_{\text{off}}/k_{\text{on}}$$
$$\propto \tau_{\text{on}}/\tau_{\text{off}}$$
$$\sim t_U^{(B)}/t_B^{(U)}$$



Swinney, D. C. Curr. Opin. Drug Discov. Devel. 12, 31–39 (2009).

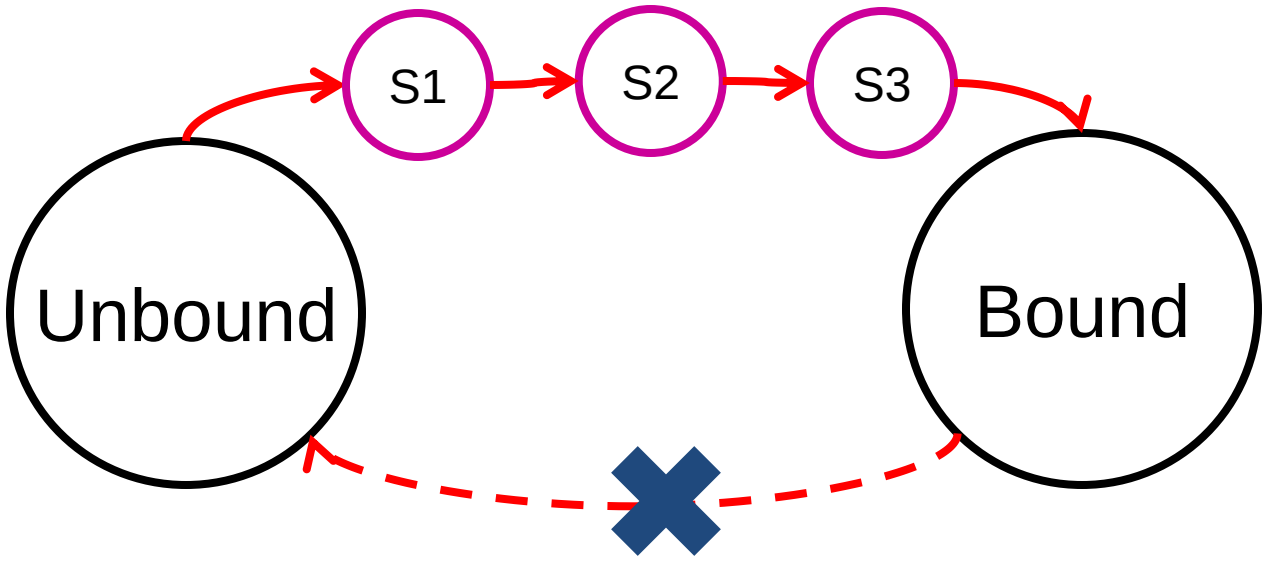
Let's move beyond this model…

On the sampled timescales…



Idea

- We introduce several *intermediate steps*

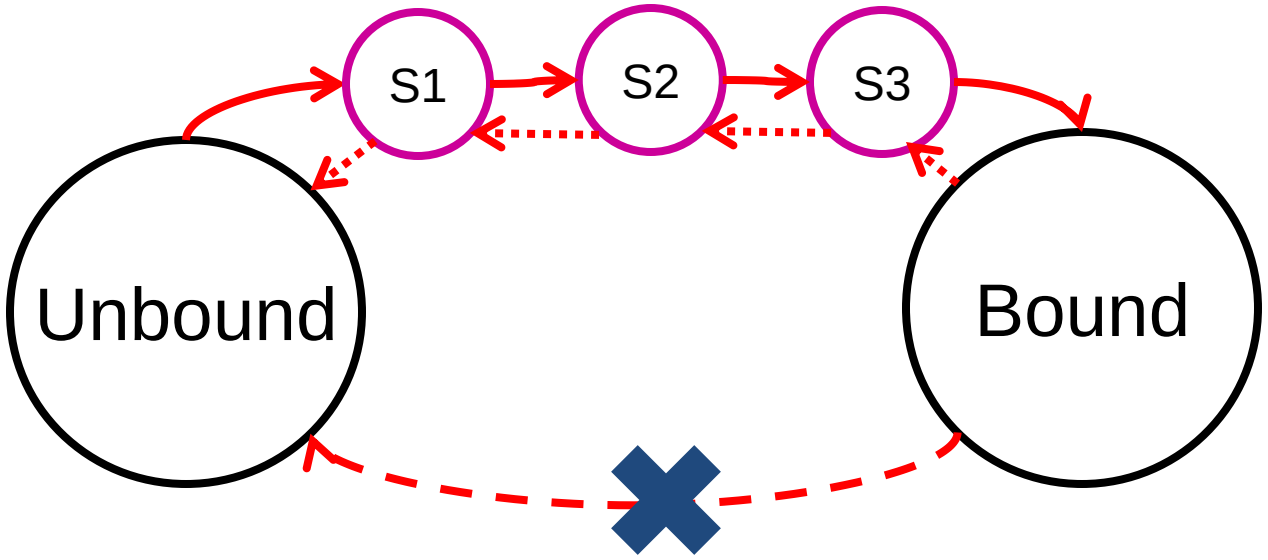


k_{on} reconstructed combining rates of different intermediate steps

Buch et al., PNAS 2012

Idea

- We introduce several *intermediate steps*
- Compute MFPT



k_{on} reconstructed combining rates of different intermediate steps

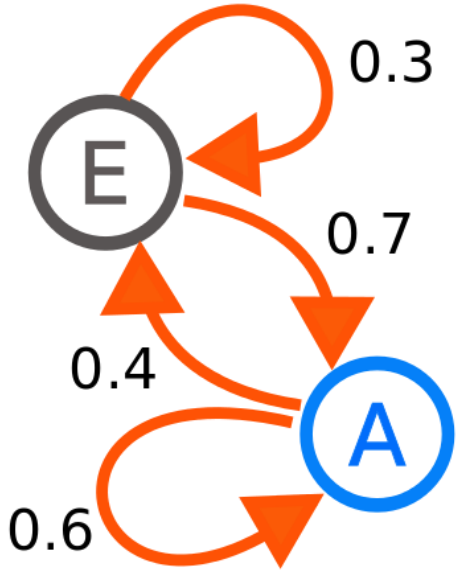
k_{off} may also be reconstructed, if states are fine-grained!

Going multi-state: Discrete-time Markov Chains

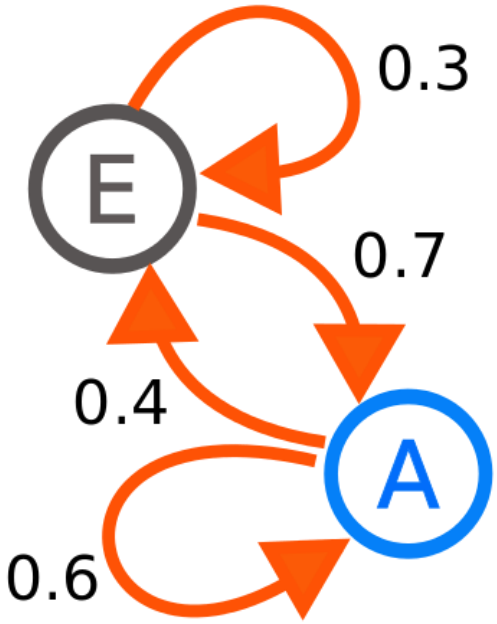
Andrei Markov
1856-1922

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

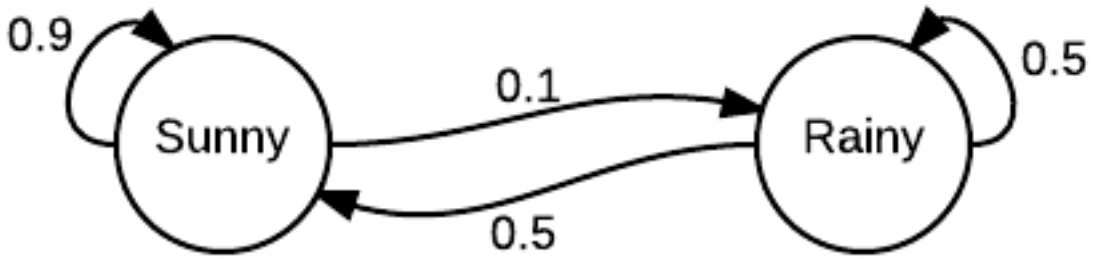


P_{ij} = *Probability to change state from i to j at each time point*

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

i \ *j*	**A**	**E**
A	0.6	0.4
E	0.7	0.3

First example



$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

Assume a deterministic initial condition:

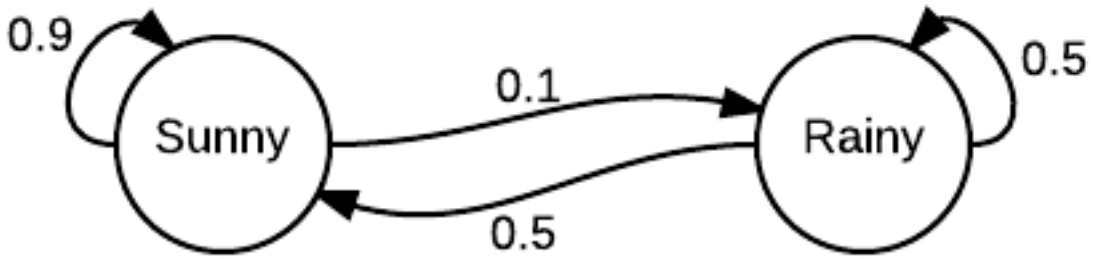
X_0 = Sunny with certainty; *i.e.*,

p(Sunny | t=0)=1 and
p(Rainy | t=0)=0 *i.e.*

$s_0 = [1, 0]$

…now, what is s_1?

First example



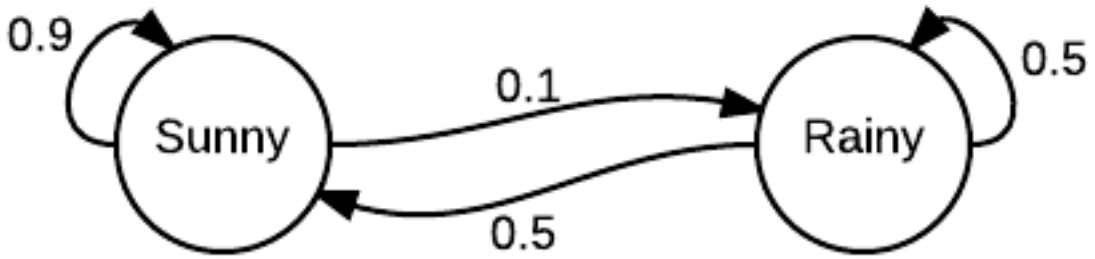
$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

What is s_1?

S_1 { p(Sunny | t=1) = 0.9
p(Rain | t=1) = 0.1

…now, what is s_2?

First example



$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

What is s_2?

s_2 {
p(Sunny | t=2) = 0.9 p(Sunny | t=1) + 0.5 p(Rainy | t=1) = 0.86
p(Rainy | t=2) = 0.1 p(Sunny | t=1) + 0.5 p(Rainy | t=1) = 0.14

In matrix form…

$\mathbf{s}_2 = \mathbf{s}_1 P$

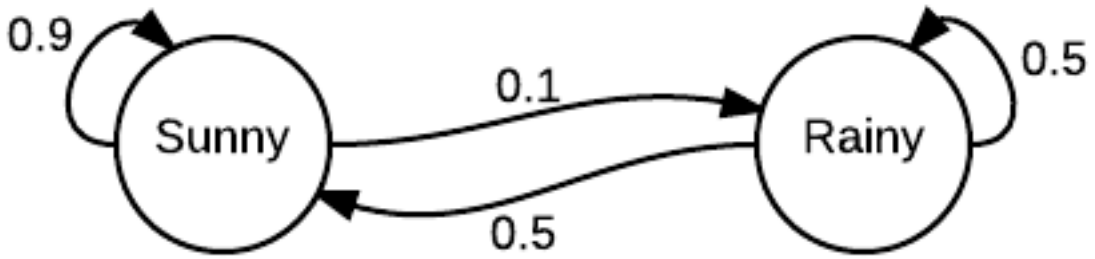
And in general…

$\mathbf{s}_{t+1} = \mathbf{s}_t P$

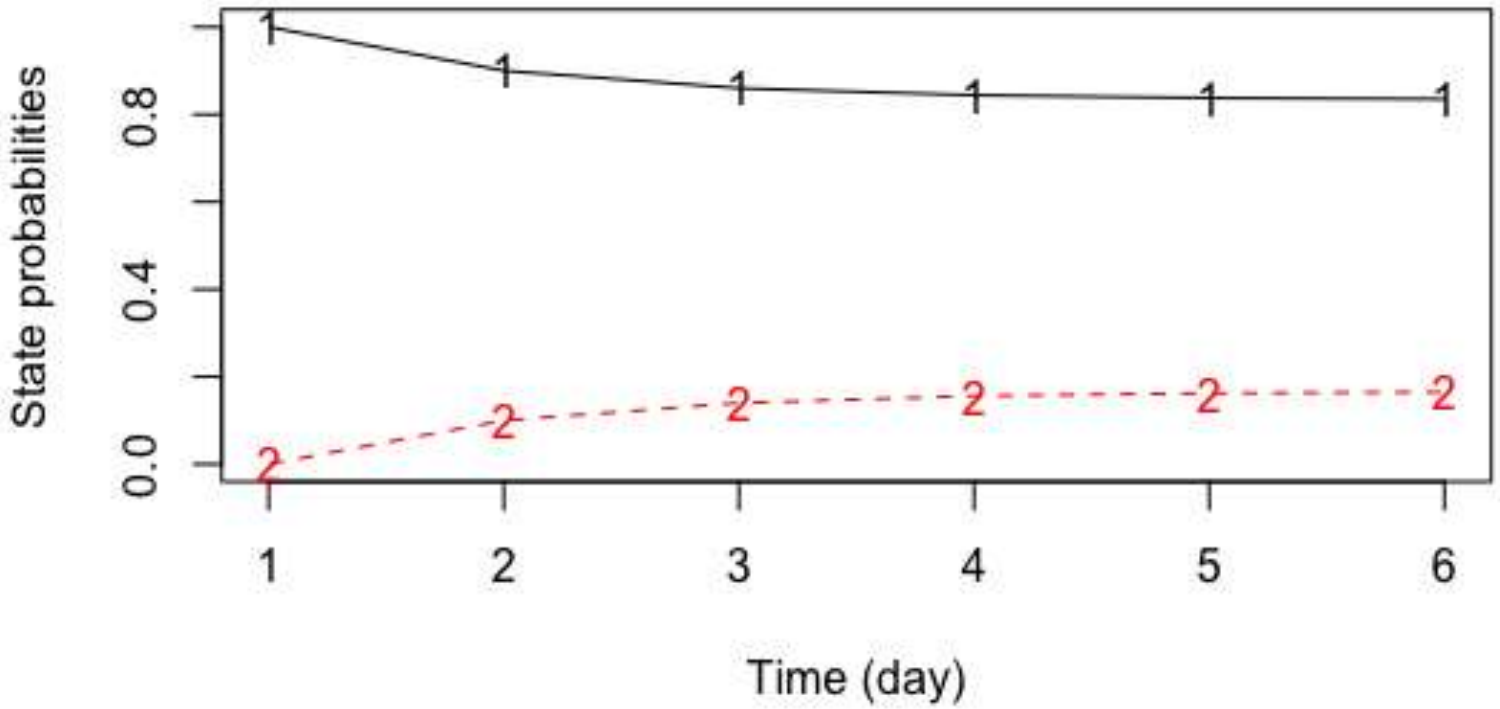
Meaning…

$\mathbf{s}_t = \mathbf{s}_0 P^t$

Let's do a numerical test…

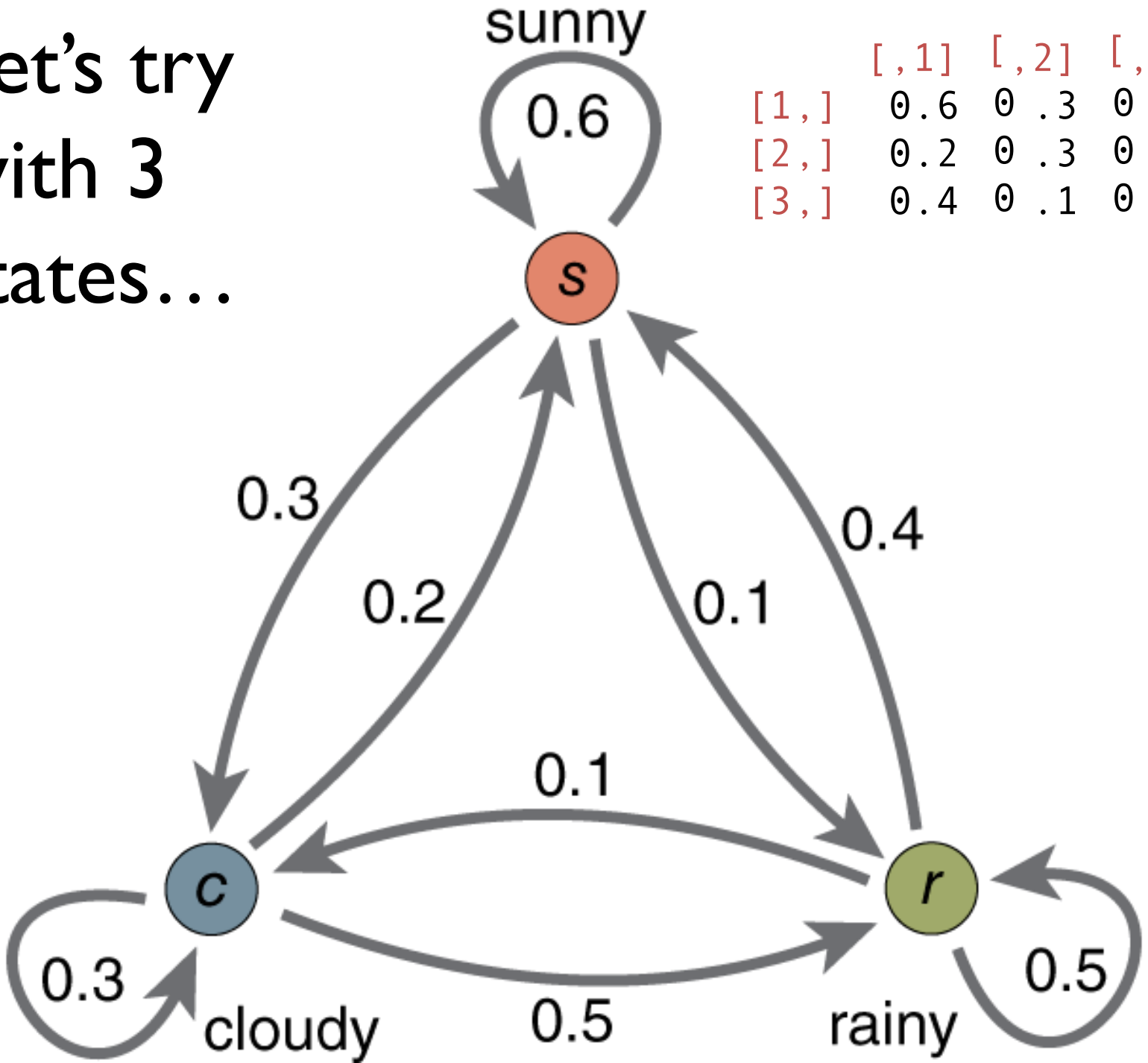


$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$



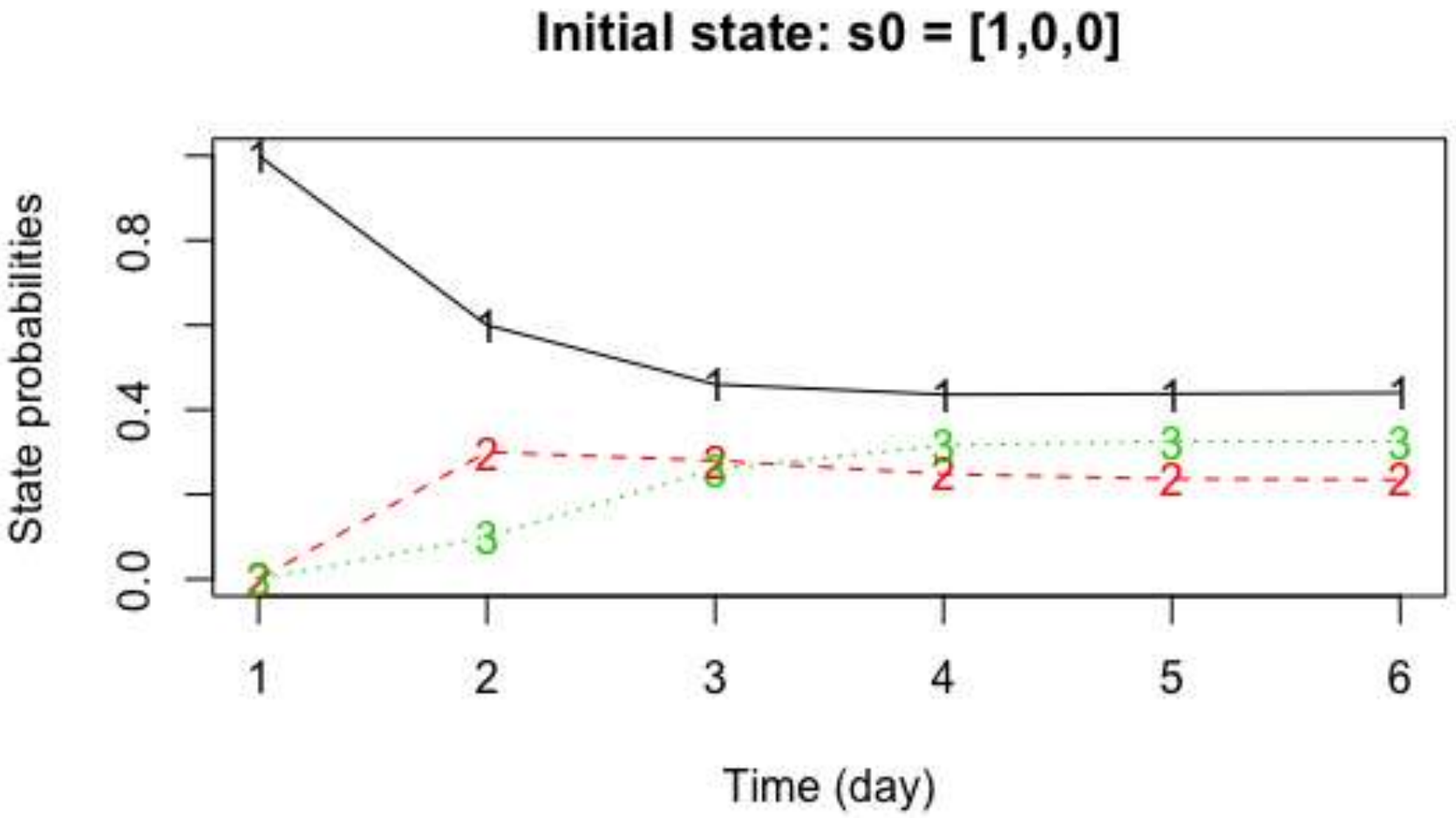
Let's try with 3 states…

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



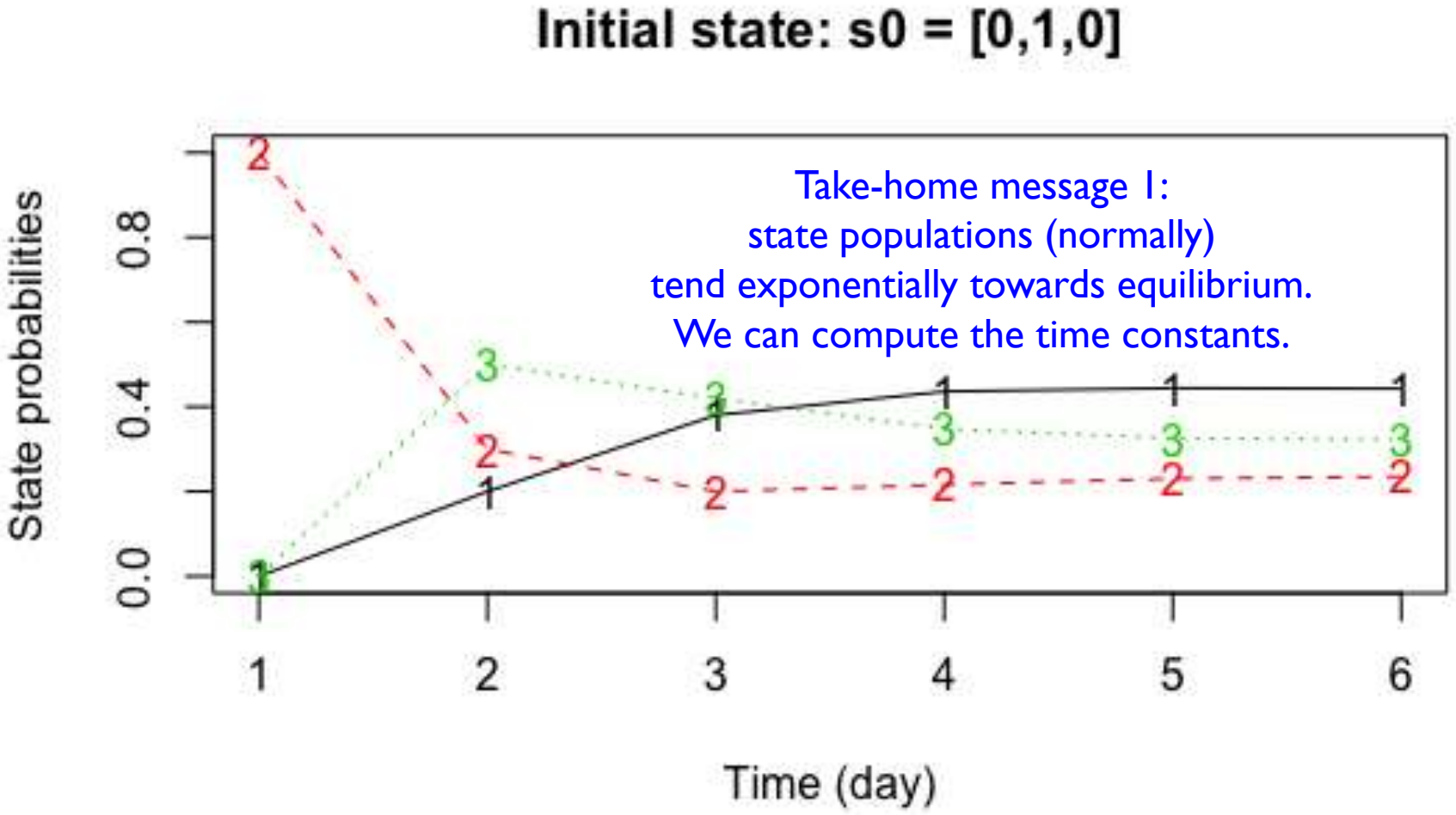
Starting from *Sunny*…


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



Starting from *Cloudy*…


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



A couple of definitions

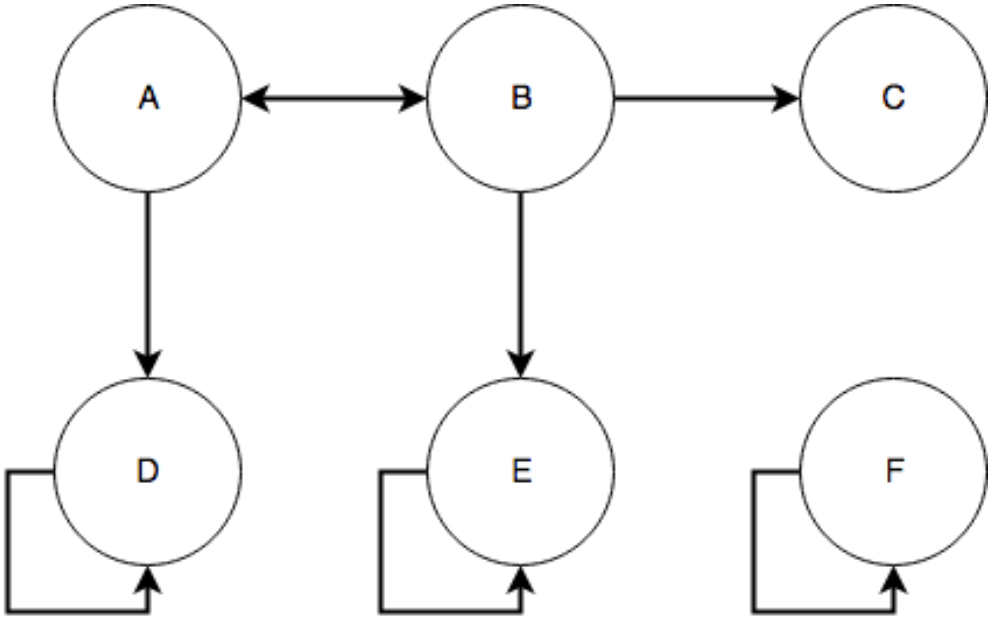
- State *j* is **accessible** from *i* (*i* → *j*) if the system has a probability ≠ 0 of going from one to the other
- *j* **communicates** with *i* iff *i* → *j* and *j* → *i*
- A set of states C is a **communicating class** if every pair of states in C communicates with each other.
- A Markov chain is said to be **irreducible** if its state space is a single communicating class.
- An **absorbing state,** once entered, cannot be left.
- (… and many more)

Homogeneity

- If the transition probabilities do not change with time, we have an *homogeneous* Markov chain
- Example of non-homogeneous MC: weather transition probabilities Markovian, but dependent on the season.

Absorbing states

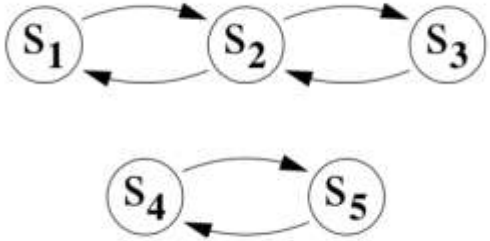
- A state is called **absorbing** if it is impossible to leave it.
- In other words, $p_{jj}=1$ (implying $p_{ij}=0$ for $i \neq j$)



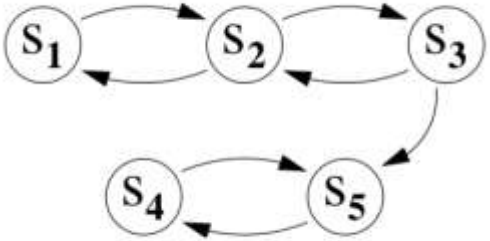
A, B are recurrent
C, D, E, F are absorbing

Irreducibility

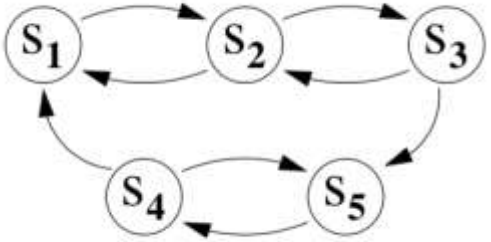
- A Markov chain is said to be **irreducible** if it is possible to get from any state to any state



Not irreducible



Not irreducible



Irreducible

Fredrik Lundin's lecture notes at
www.math.chalmers.se/Stat/Grundutb/Chalmers/TMS081/oldpage/Lecture_notes/lecture3.pdf

Important quantities we can compute

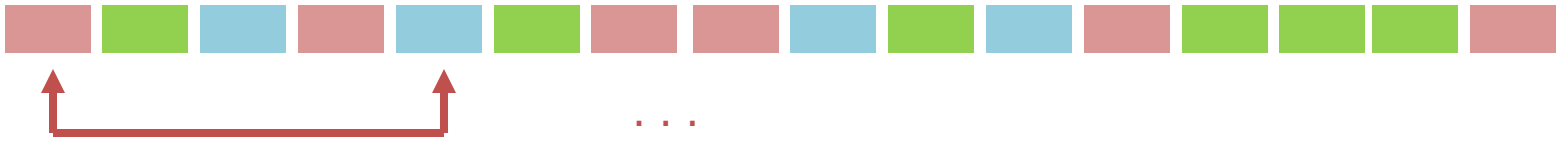
- Stationary distribution (→ eq. probabilities)
- Relaxation times
- Mean-first passage times (→ kinetic rates)
- And others we won't discuss:
  - Committor probabilities
  - Fluxes
  - …

Learning transition matrices from trajectories (of discrete states)

Markov models

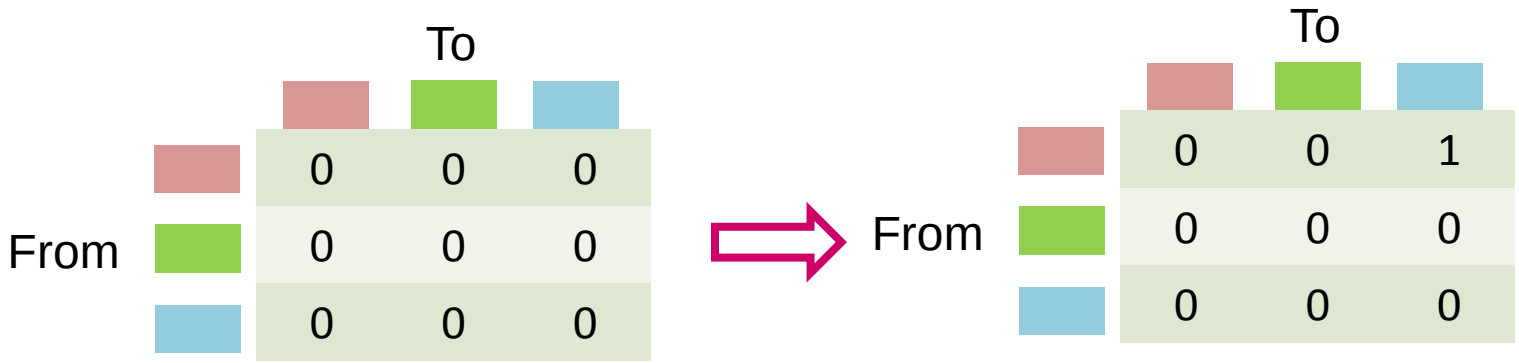
Total sampled time (e.g. 100 μs)

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



Lag time τ (here: 4 time units)

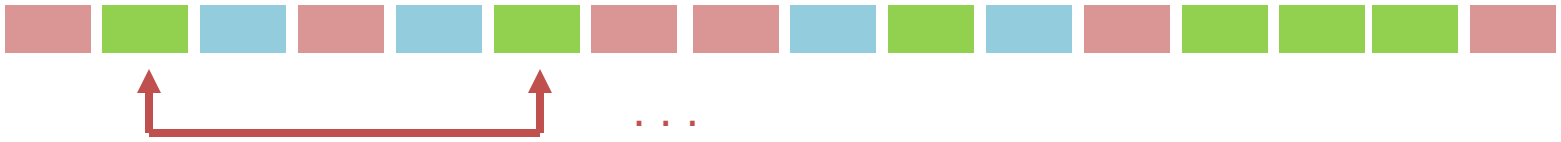
Compute the transition probability matrix sliding a window of lag time τ



Markov models

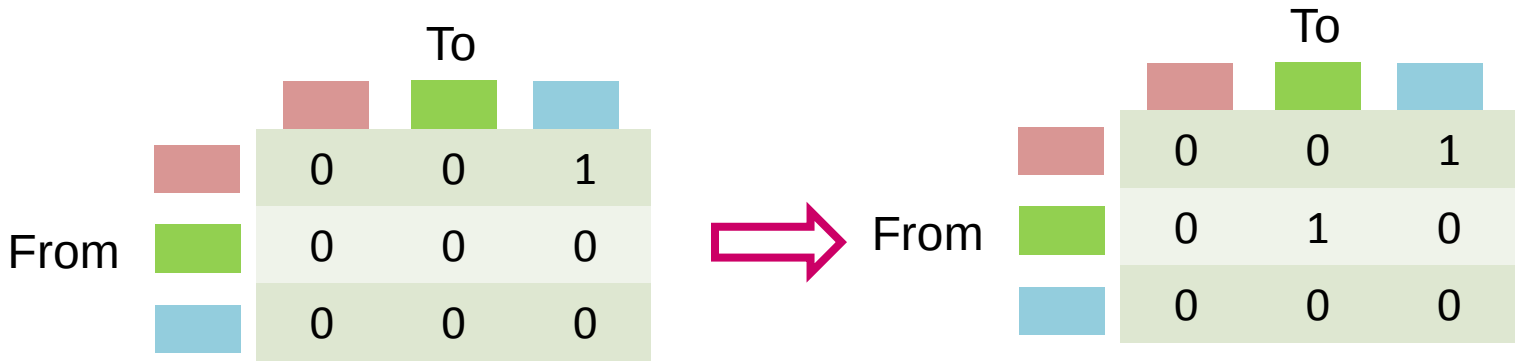
Total sampled time (e.g. 100 μs)

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



Lag time τ (here: 4 time units)

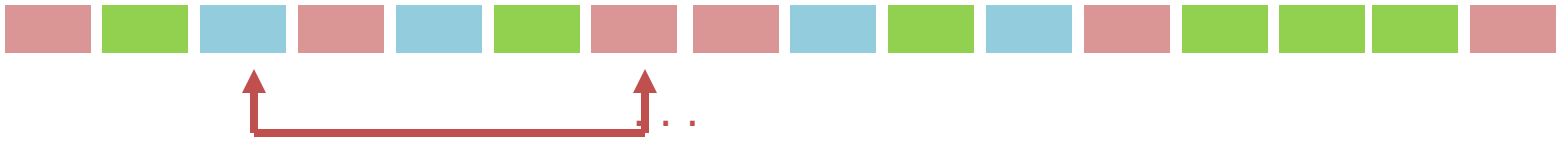
Compute the transition probability matrix sliding a window of lag time τ



Markov models

Total sampled time (e.g. 100 μs)

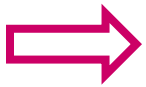
Define the *discrete* state of the system in *discrete time* (e.g. via reaction coordinates)



Lag time τ (here: 4 time units)

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

From \ To			
	0	0	1
	0	1	0
	0	0	0

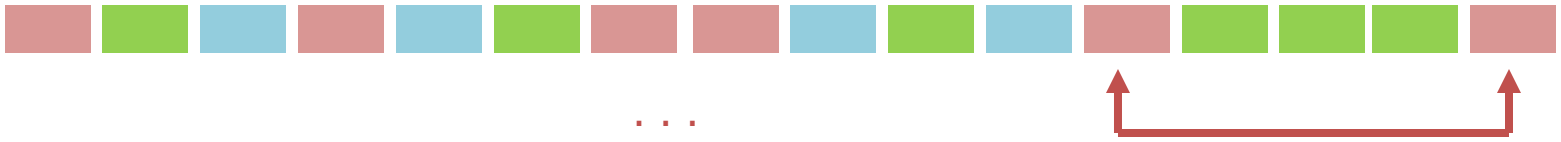


From \ To			
	0	0	1
	0	1	0
	1	0	0

Markov models

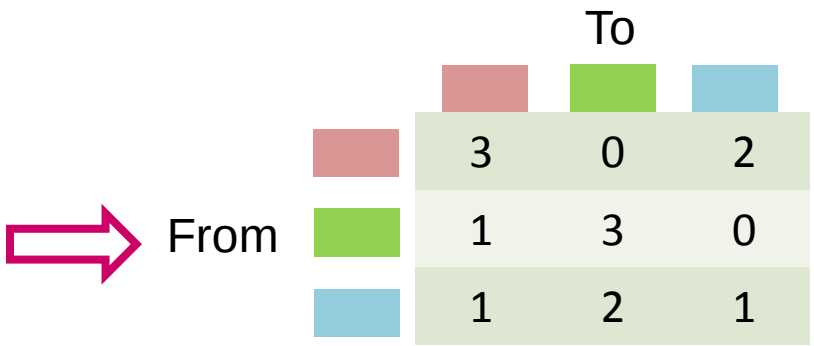
Total sampled time (e.g. 100 μs)

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



Repeat until the end of the trajectory.

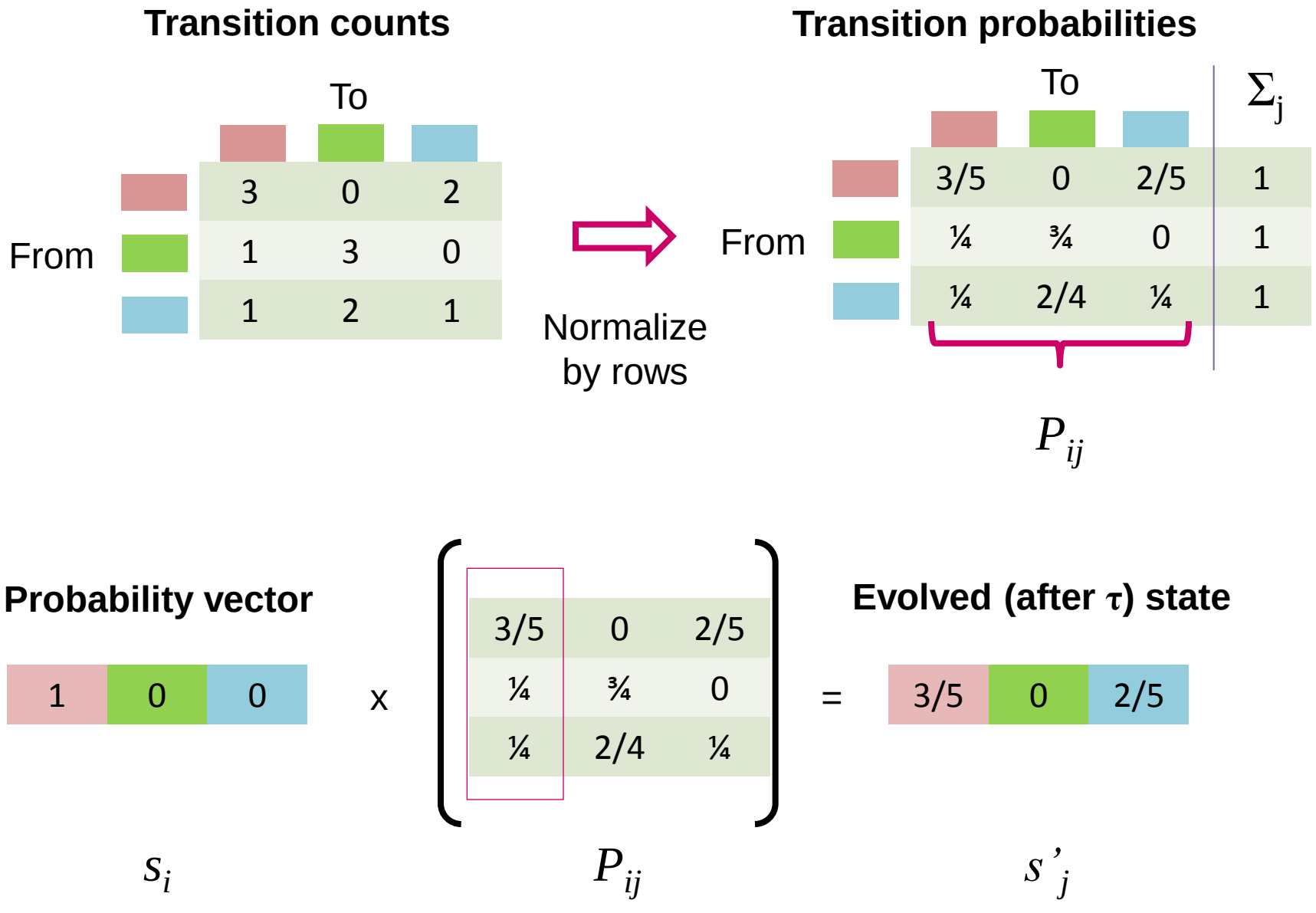
Note the Markovian assumption: transition probabilities **do not** depend on history (neither, for us, on time).

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

Important: you can accumulate counts
from **different** trajectories in the **same** matrix!

It's all valid, independent sampling
of the same underlying system.

Example: multiple weather stations in the same city.
They may be active at different times!


Transition counts
To
From
3 0 2
1 3 0
1 2 1
Normalize by rows
Transition probabilities
To
Σj
From
3/5 0 2/5 1
¼ ¾ 0 1
¼ 2/4 ¼ 1
Pij
Probability vector
1 0 0
x
3/5 0 2/5
¼ ¾ 0
¼ 2/4 ¼
=
3/5 0 2/5
si
Pij
s'j

Probability vector

1	0	0

s_i

x

$$\begin{pmatrix} 3/5 & 0 & 2/5 \\ ¼ & ¾ & 0 \\ ¼ & 2/4 & ¼ \end{pmatrix}$$

P_{ij}

=

Evolved state

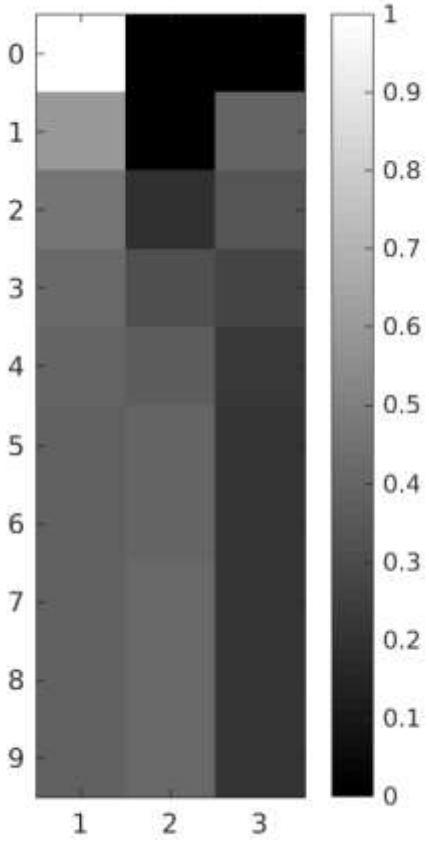
3/5	0	2/5

s'_j

$$s' = sP$$
$$s'' = (sP)P = sP^2$$
$$s^{(n)} = sP^n$$

1	0	0
.60	0	.40
.46	.20	.34
.41	.32	.27
.39	.37	.23
.39	.40	.22
…	…	…

time



Probability vector

1	0	0

s_i

x

3/5	0	2/5
¼	¾	0
¼	2/4	¼

P_{ij}

=

Evolved state (after tau)

3/5	0	2/5

s'_j

$$s^{\infty} P = s^{\infty}$$

$$s' = sP$$
$$s'' = (sP)P = sP^2$$
$$s^{(n)} = sP^n$$
$$\ldots$$
$$s^{\infty} = \quad ?$$

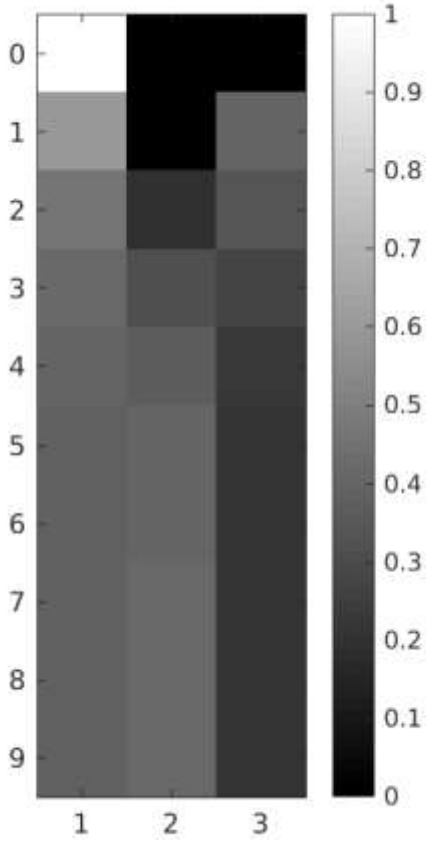
Left eigenvector of P (eigenvalue 1)

Is the stationary state
= equilibrium probabilities
= the free energy surface

```
[a,b]=eig(P')
a(:,3)/sum(a(:,3))
= 0.385 0.410 0.205 = [5/13 16/39 8/39]
```

Relaxation towards equilibrium

time



Equilibrium is reached within typical **relaxation times***.

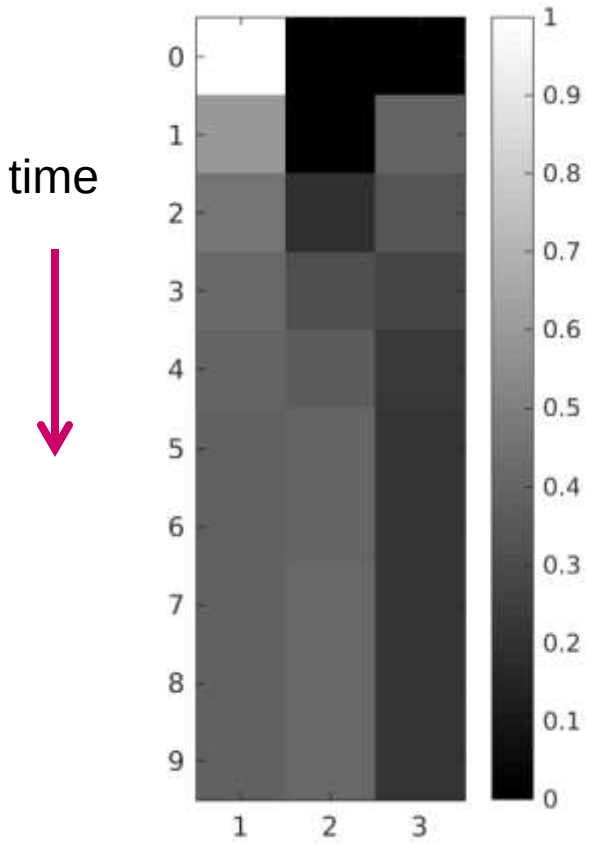
Do an eigen-decomposition…

$$\mathbf{x}VP = \Lambda\mathbf{x}V$$
$$\ldots \text{with } \Lambda = \text{diag}(\mu_i)$$
$$\mathbf{y} \equiv \mathbf{x}V$$
$$\mathbf{y}(t+1) = \Lambda\mathbf{y}(t)$$
$$y_i(t) = y_i(0)\mu_i^t = y_i(0)e^{t \log \mu_i}$$

*

Current Opinion in Structural Biology 2014, 25:135–144

Relaxation towards equilibrium



Equilibrium is reached within typical **relaxation times**.

They are computed from the eigenvalues < 1…

$$\tau_k = -\frac{\tau}{\ln \mu_k(\tau)}$$

…and called *implied timescales* (they depend on τ)

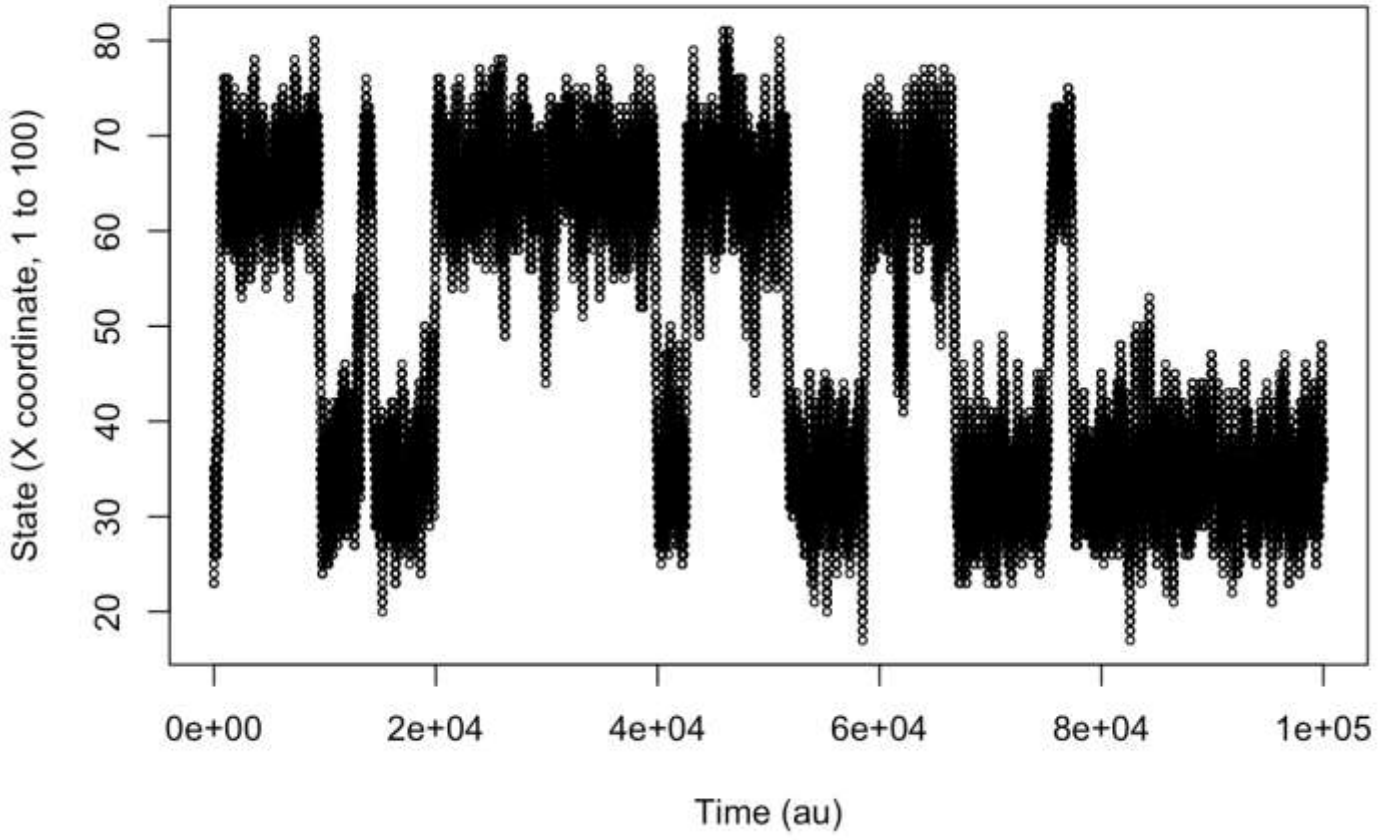
Current Opinion in Structural Biology 2014, 25:135–144

Markov modeling a 1D trajectory

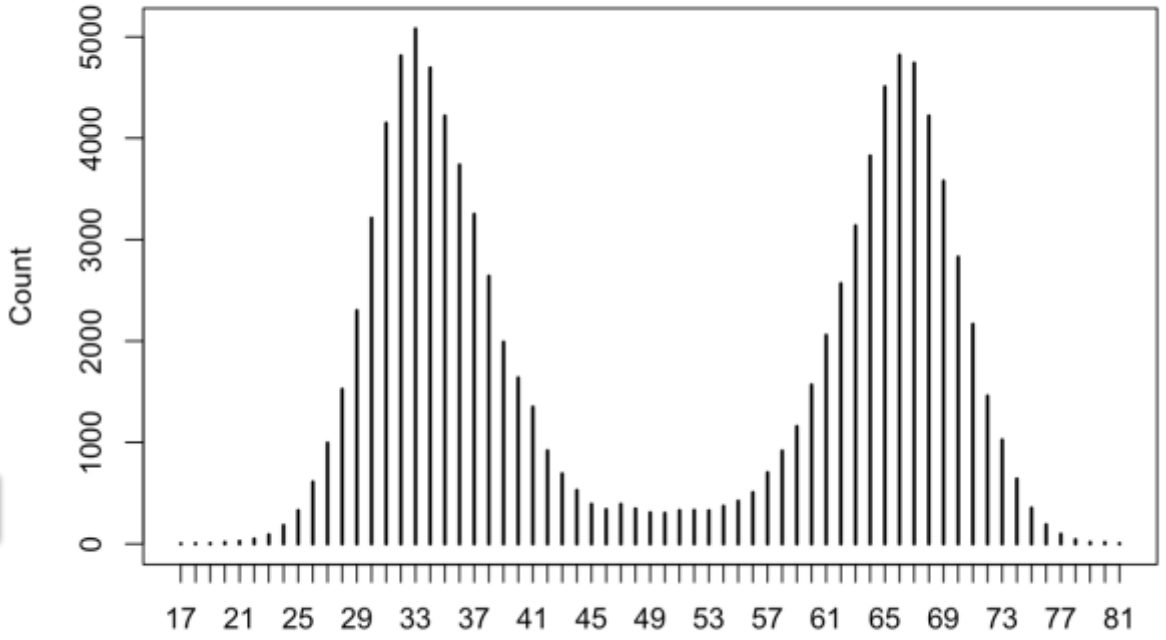
(Please find the extended version online,
"Markov state models of a 1D trajectory",
with R code)

Start with a 1-D trajectory

- Already discretized in 100 bins

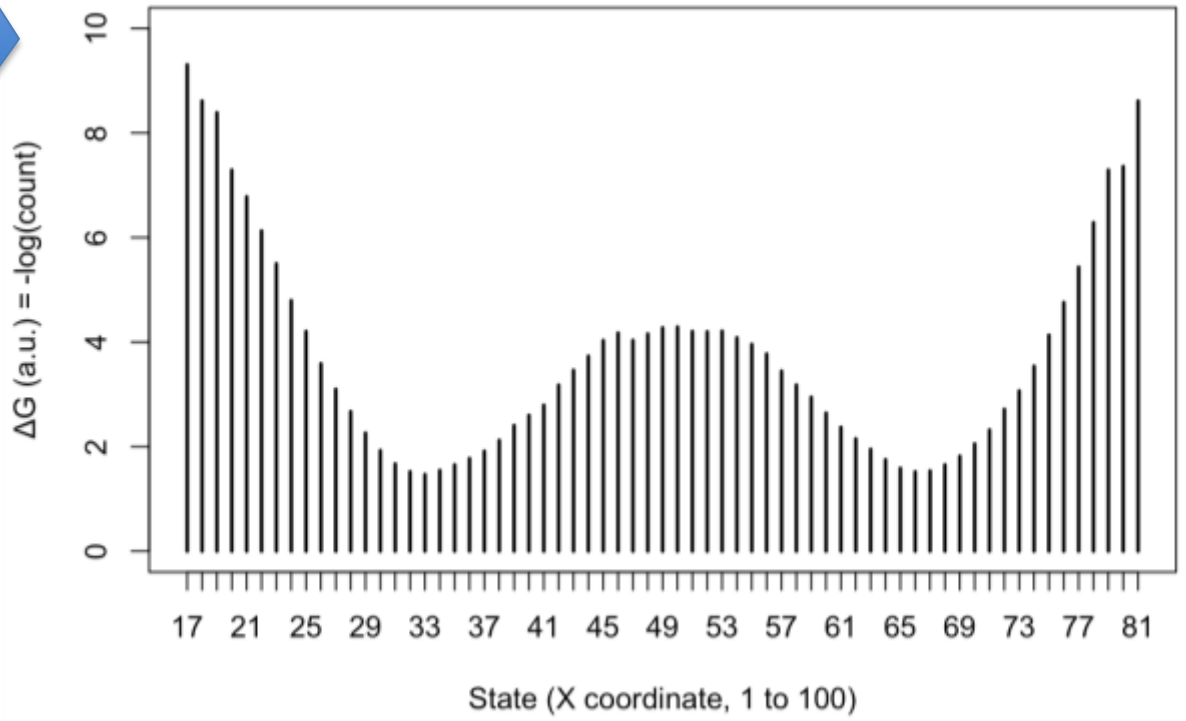


Histogram



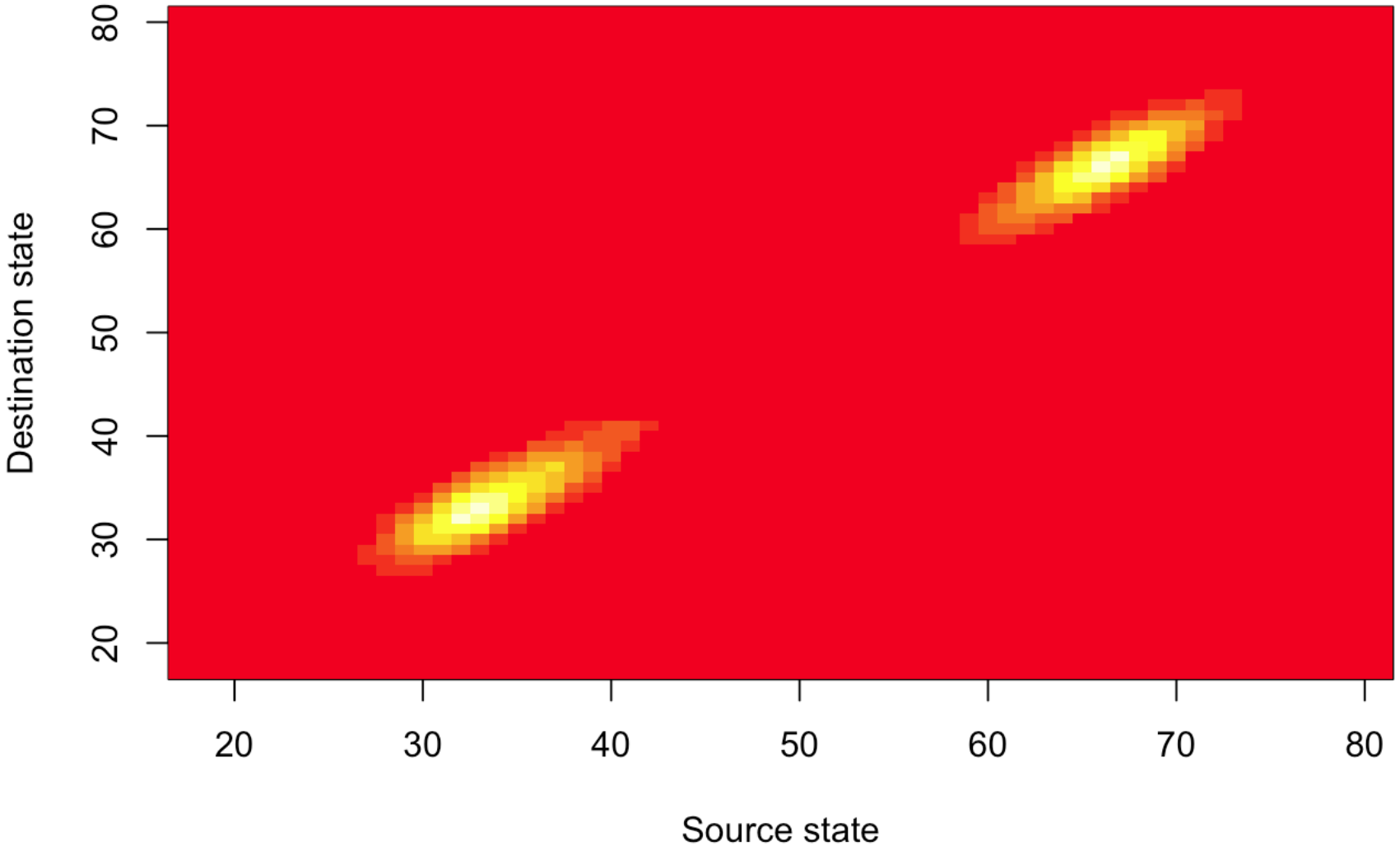
Boltzmann inversion

-log(count)



The transition *count* matrix*

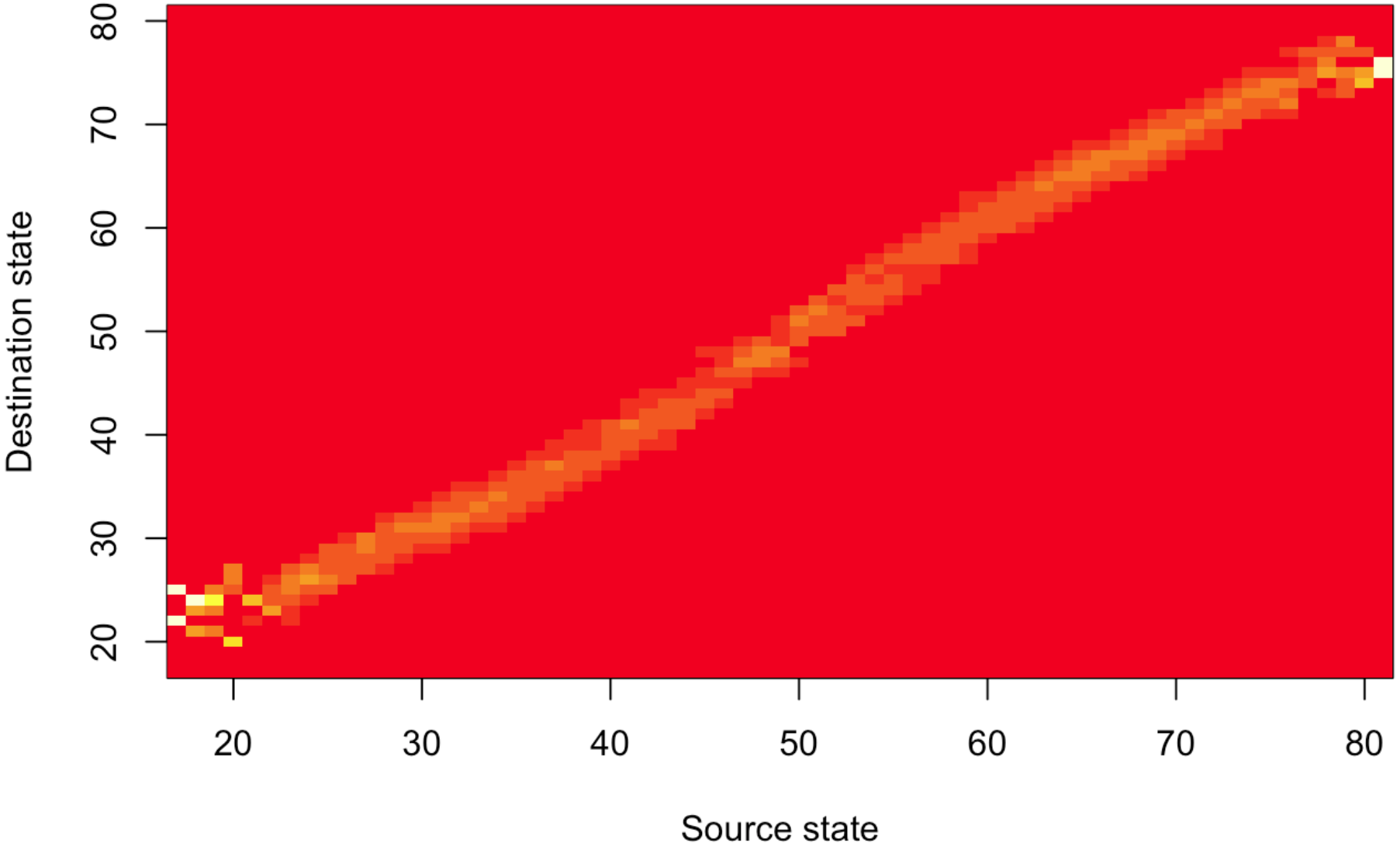
How many times we have seen state i going to j after τ=10 time units



* shown as an image for compactness

The transition *probability* matrix

Rows normalized to sum to 1

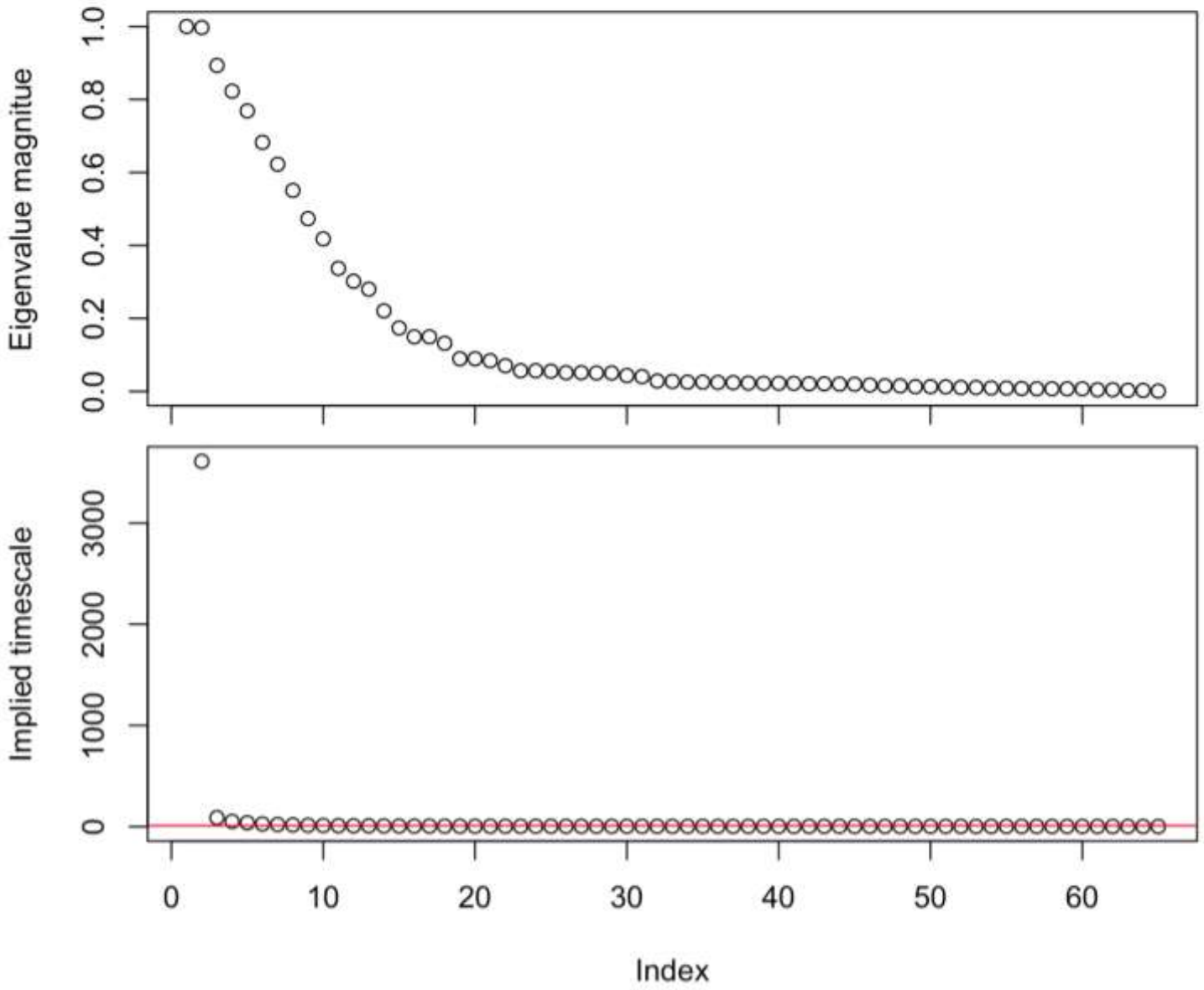


Eigenvalues (at τ=10)

Eigenvalues μ_i

Implied timescales

$-\tau/\log(\mu_i)$



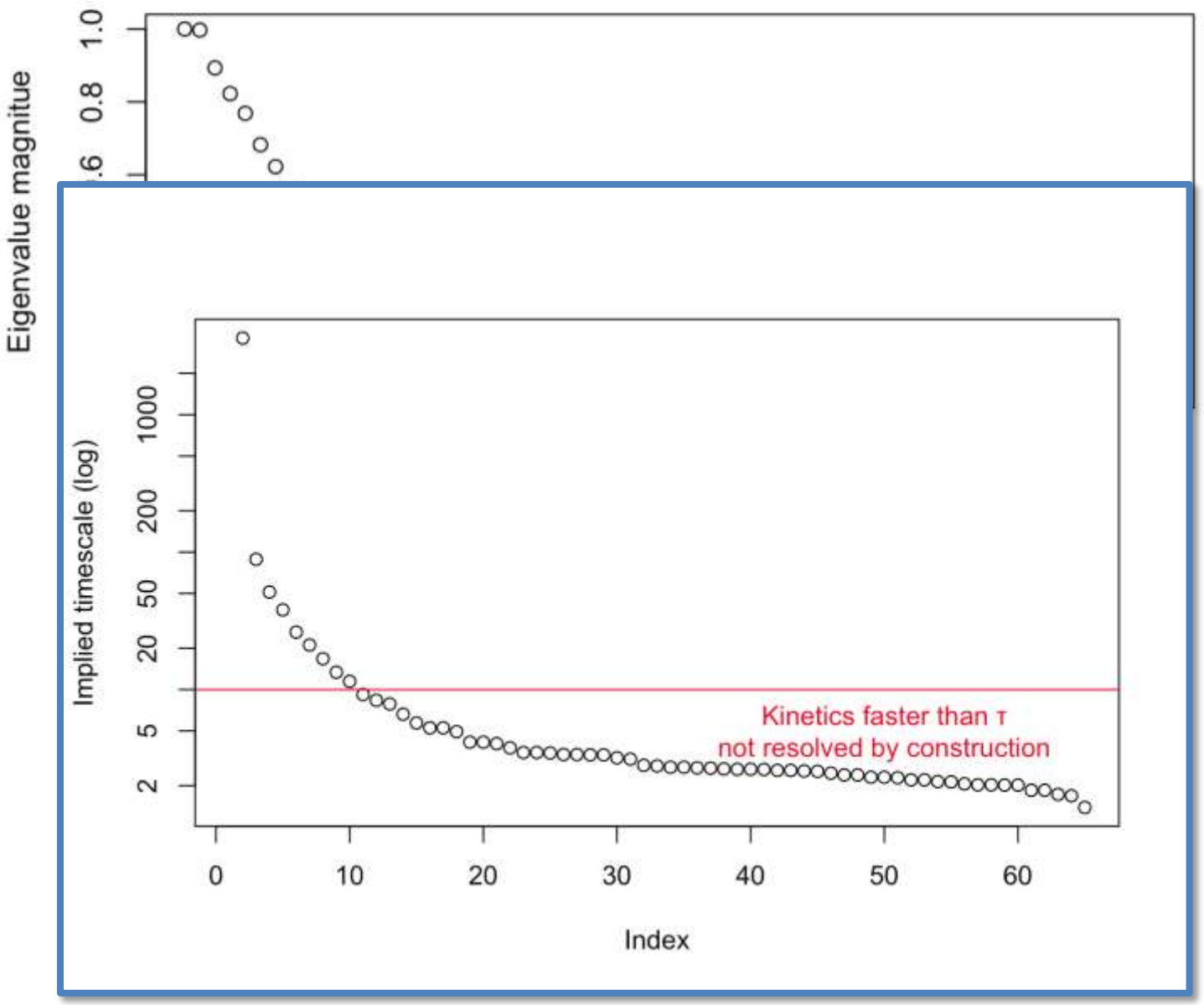
Eigenvalues (at τ=10)

Eigenvalues μ_i

Implied timescales

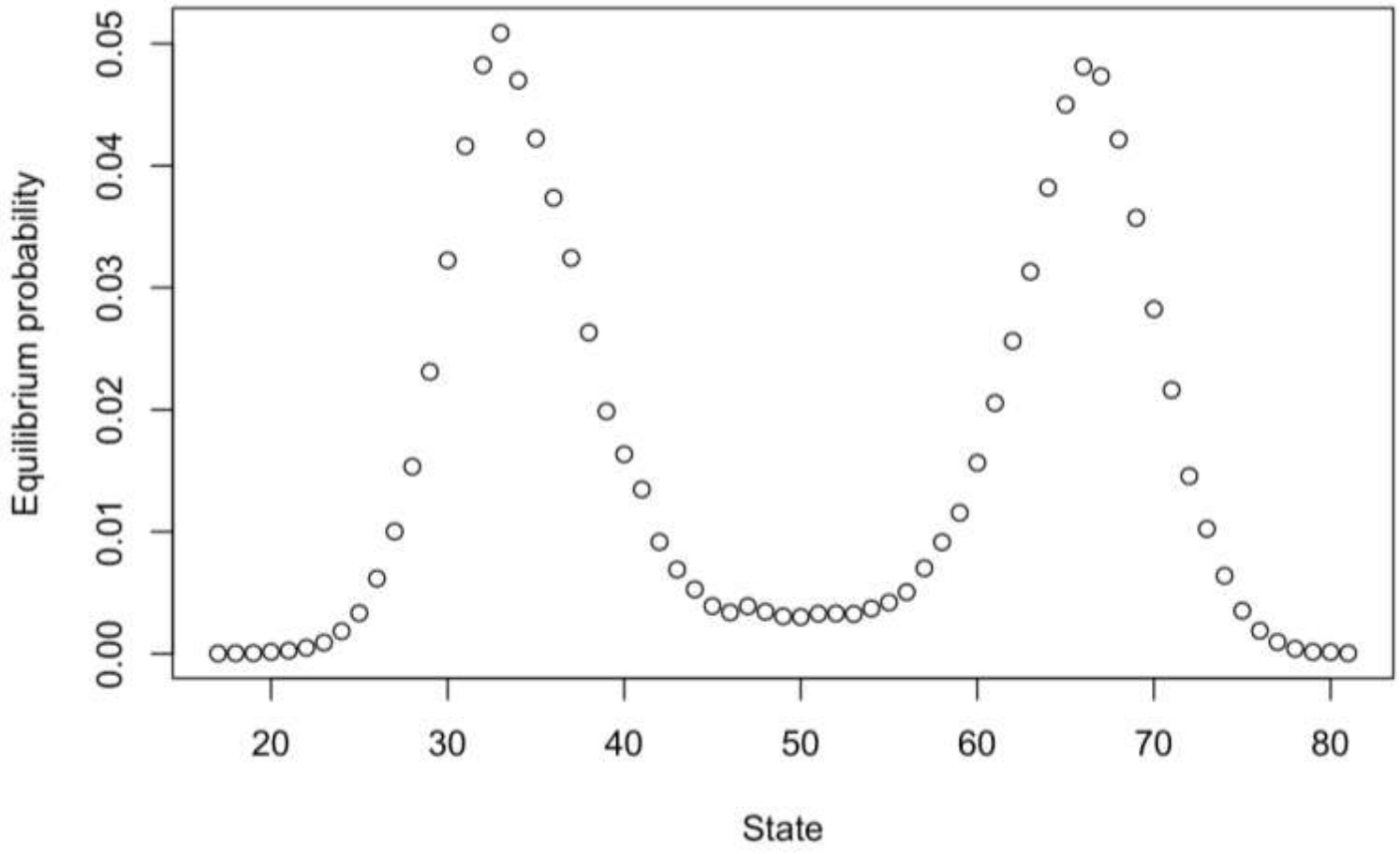
$-\tau/\log(\mu_i)$

Now in log scale!



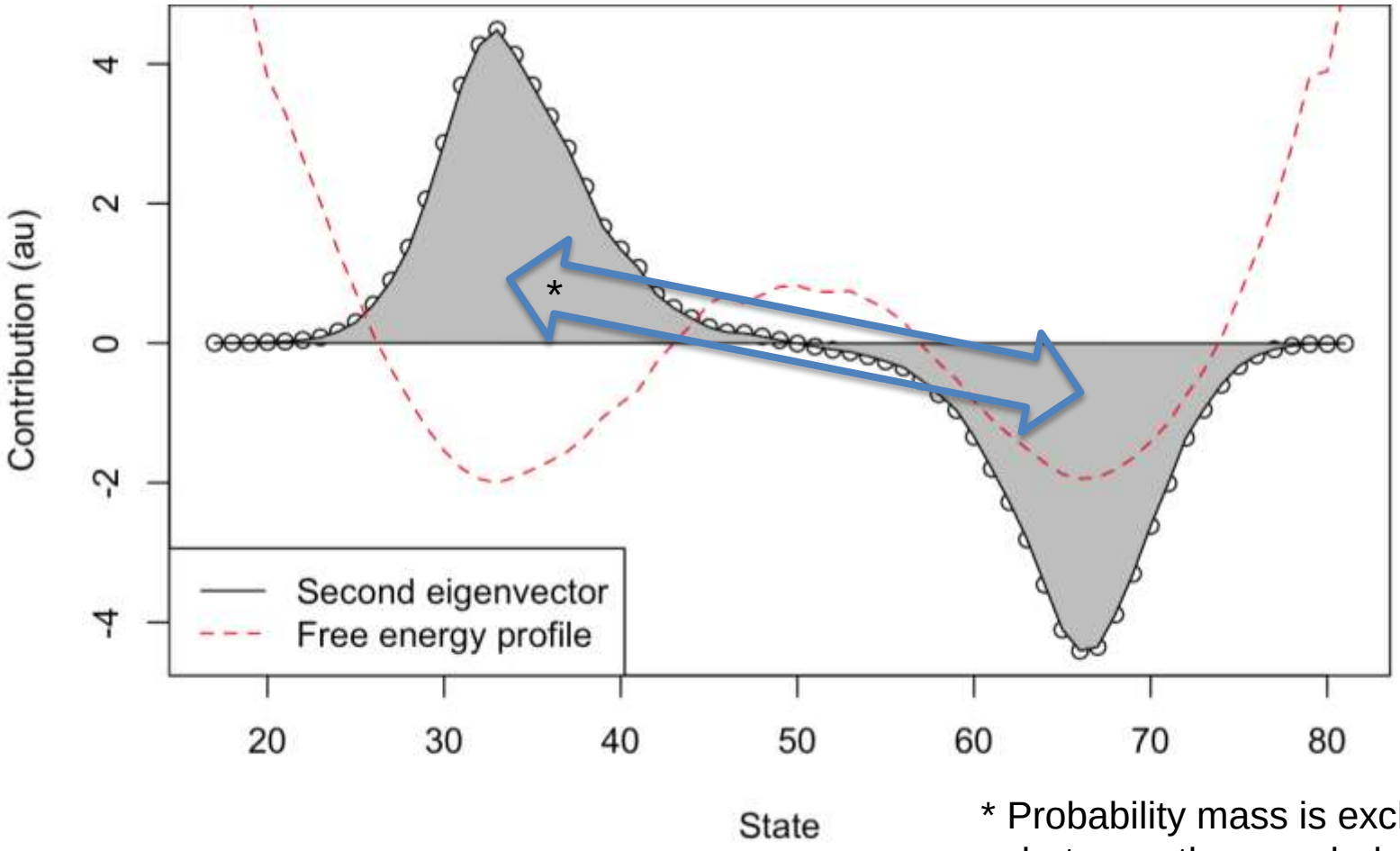
First eigenvector ($\mu_1=1$)

This is the stationary state (normalize so it sums to 1)



Second eigenvector (μ_2=0.997)

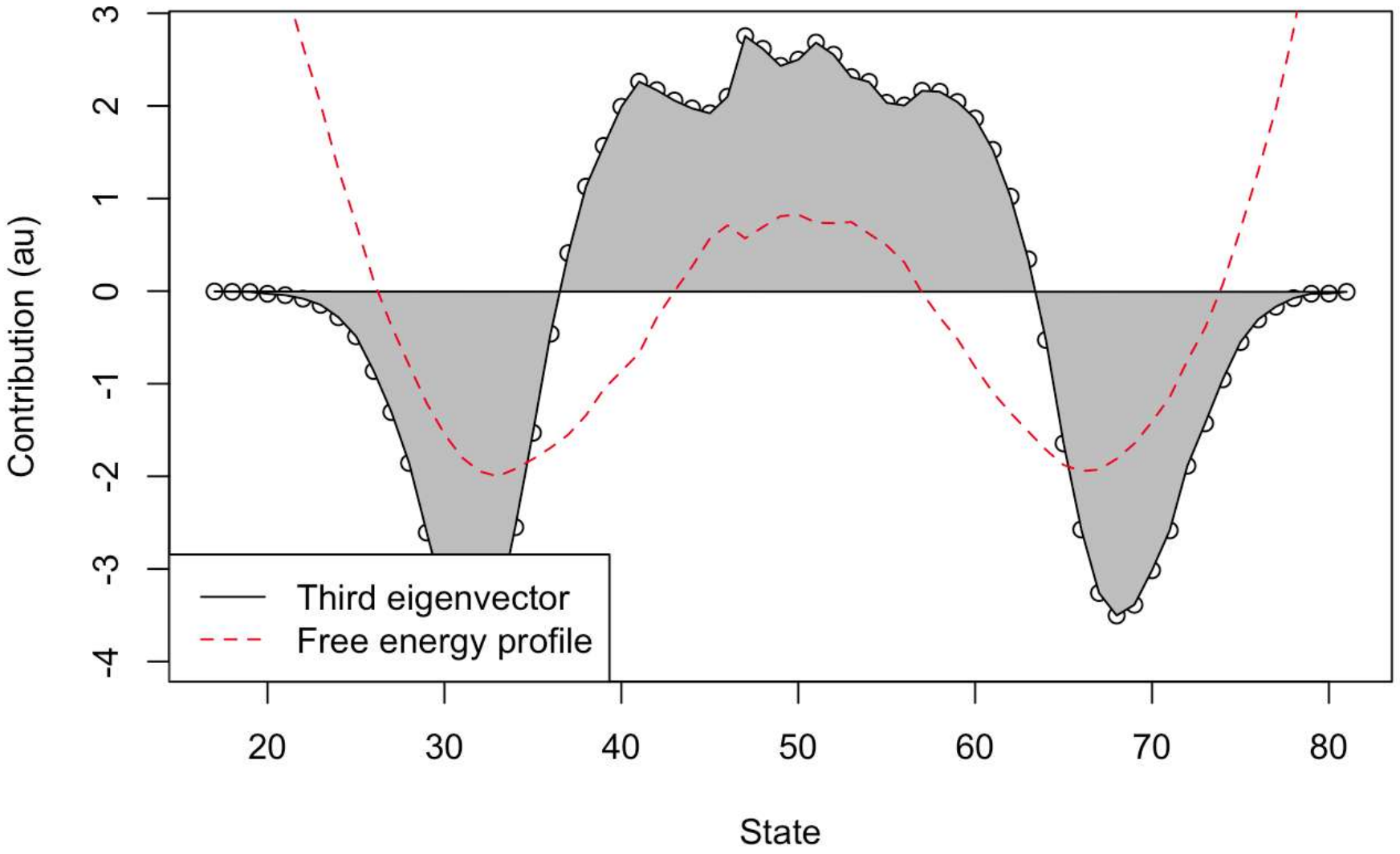
This is the slowest relaxation mode: ITS τ_2 = 3610 time units



* Probability mass is exchanged between the + and - basins

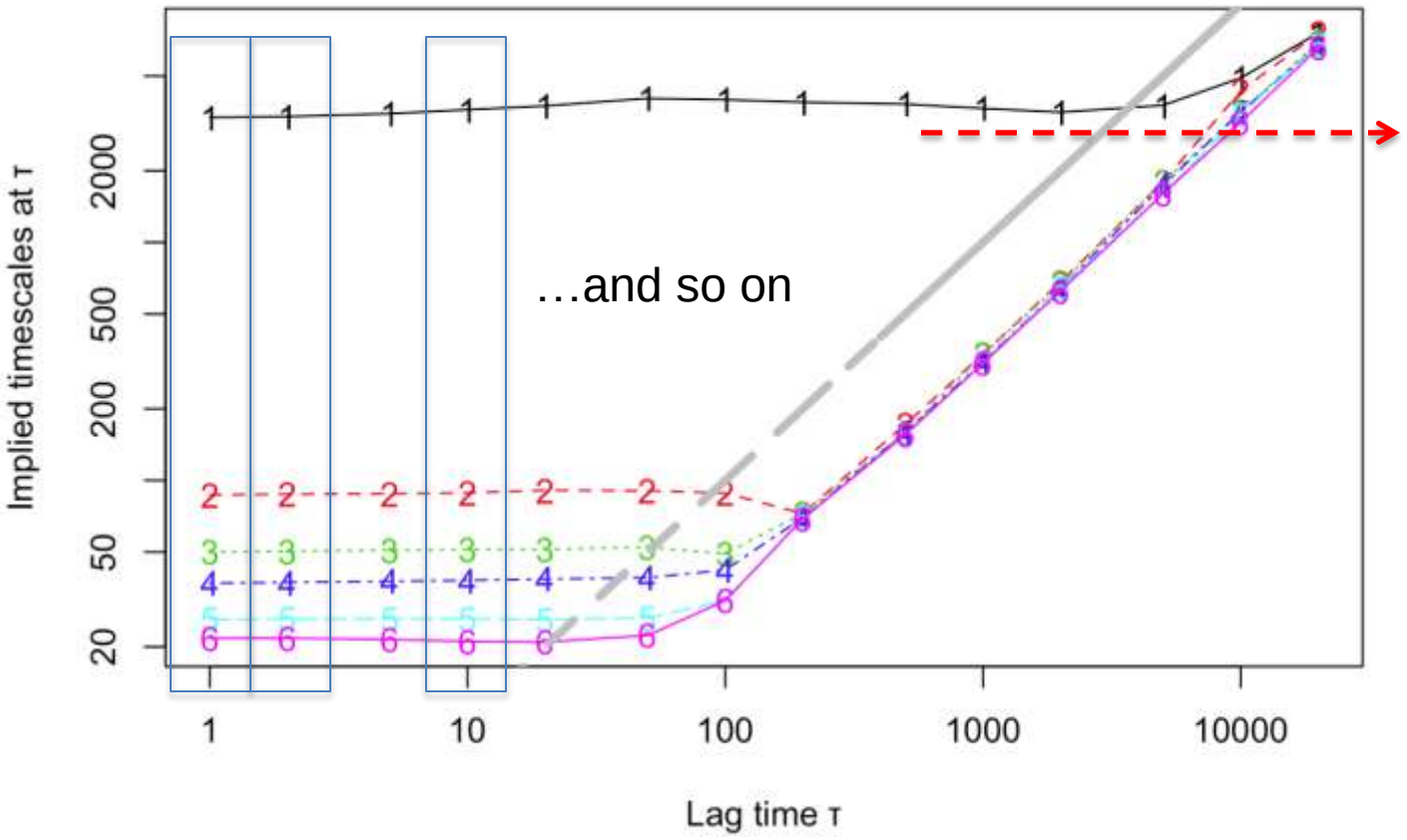
3rd eigenvector (μ_3=0.893)

This is the slowest relaxation mode: ITS τ_3 = 88.7 time units



Implied timescales plot

Repeat the eigenvalues determination for several lags. Check convergence.



Here, convergence is achieved very early.

Reasons:

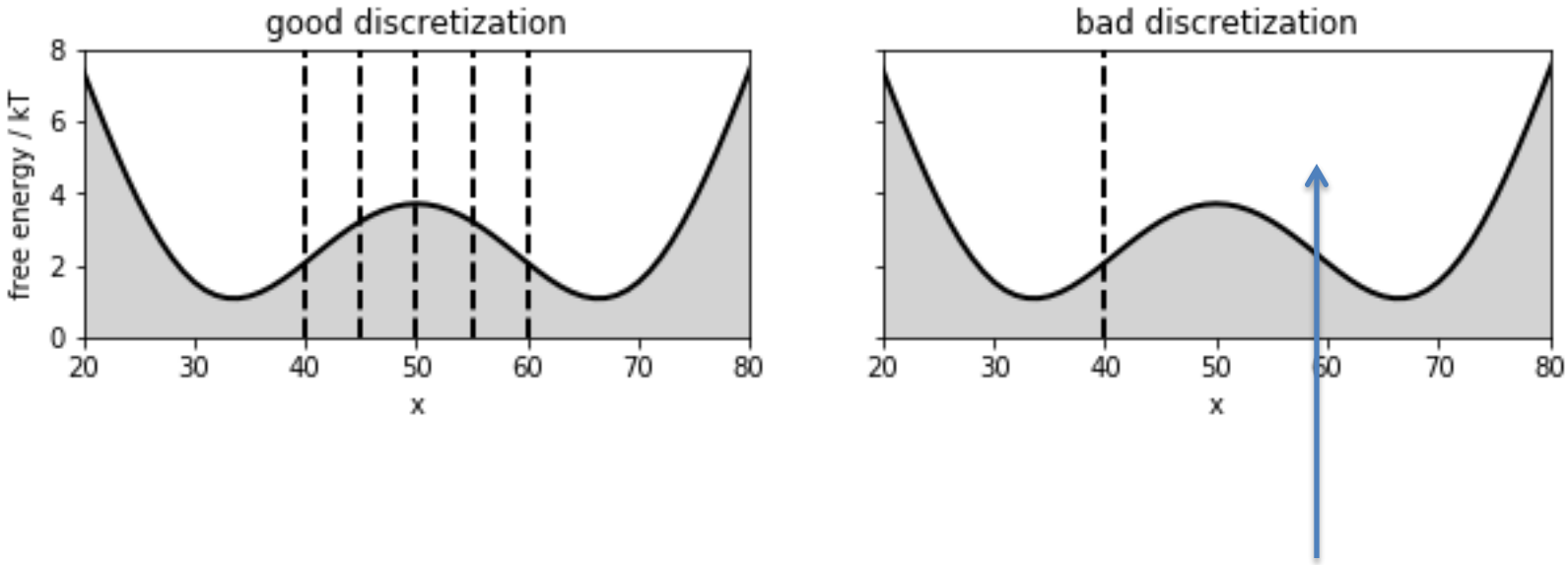
(a) true two-state dynamics;

(b) absence of orthogonal degrees of freedom;

(c) fine space discretization

Macrostates

A bad choice of the discretization breaks the Markovianity assumption

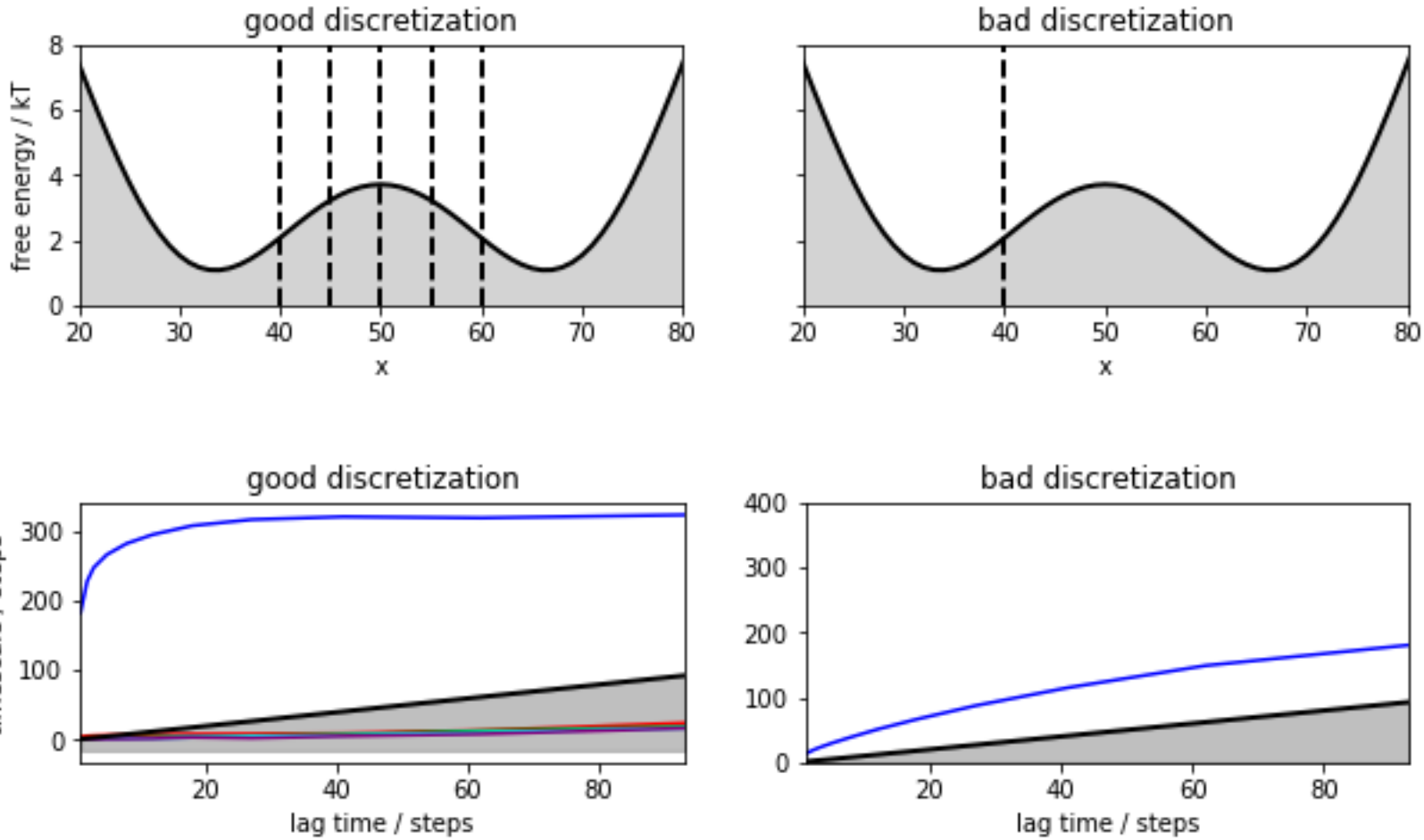


In the “bad discretization” case, the barrier is embedded in one of the states. This generates a “long term memory” effect: the rightmost state could actually be short-lived (if we are on the left of the barrier) or long-lived (if we are on its right). These two cases are convoluted into the same, so that the present state information itself is not sufficient to predict the “future” of the system any more.

Based on http://www.emma-project.org/v2.0.4/generated/model_selection_validation.html

Macrostates

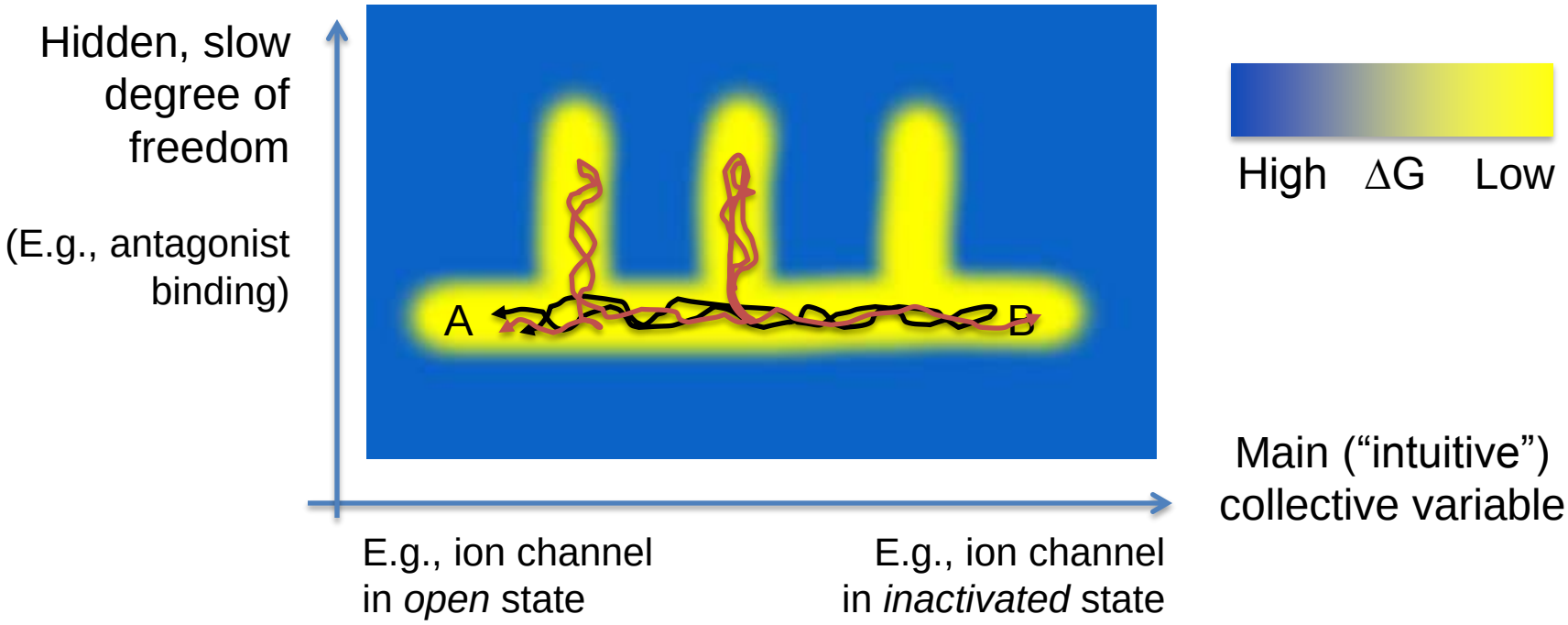
A bad choice of the discretization breaks the Markovianity assumption



Based on http://www.emma-project.org/v2.0.4/generated/model_selection_validation.html

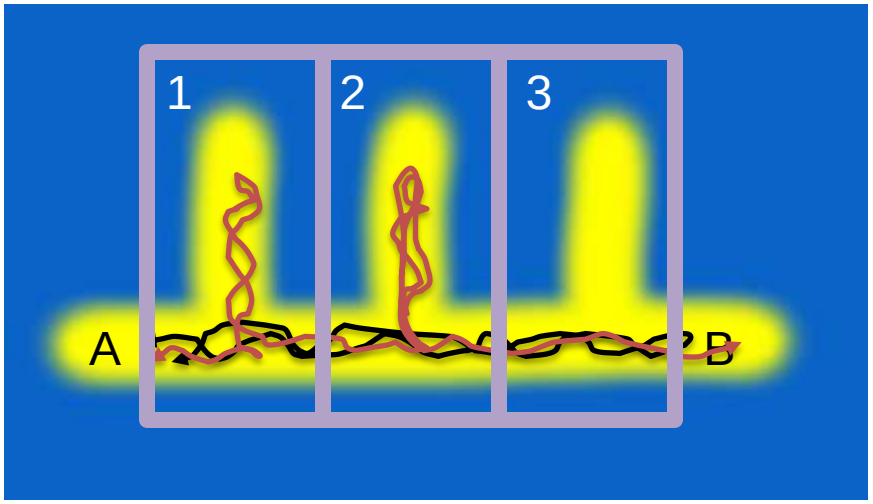
Orthogonal degrees of freedom

- Ideally, an intuitive collective degree of freedom would be a good reaction coordinate
- This is unfortunately rarely the case. E.g…

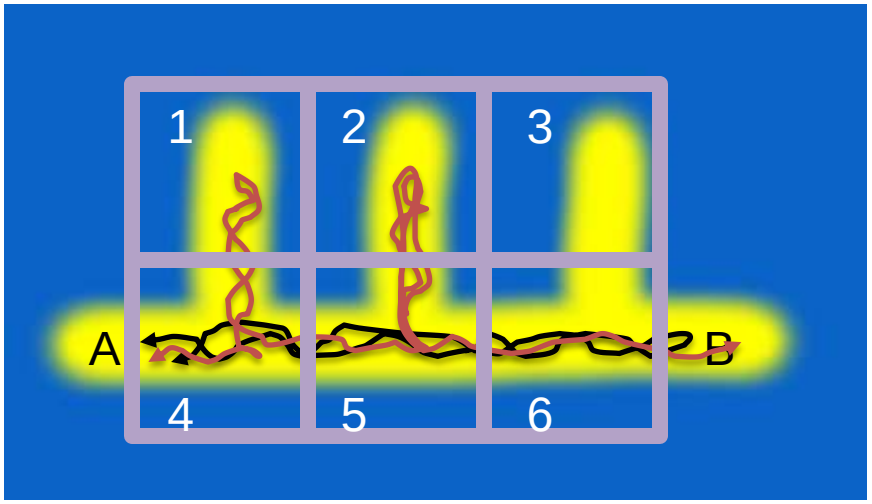


Orthogonal degrees of freedom

- Biased sampling: include bias along the "vertical" axis (i.e. make it part of the reaction coordinate)
- MSM: make distinct states
  - Otherwise, implied timescales plot will show non-convergence*



Not ok (violates Markovianity)



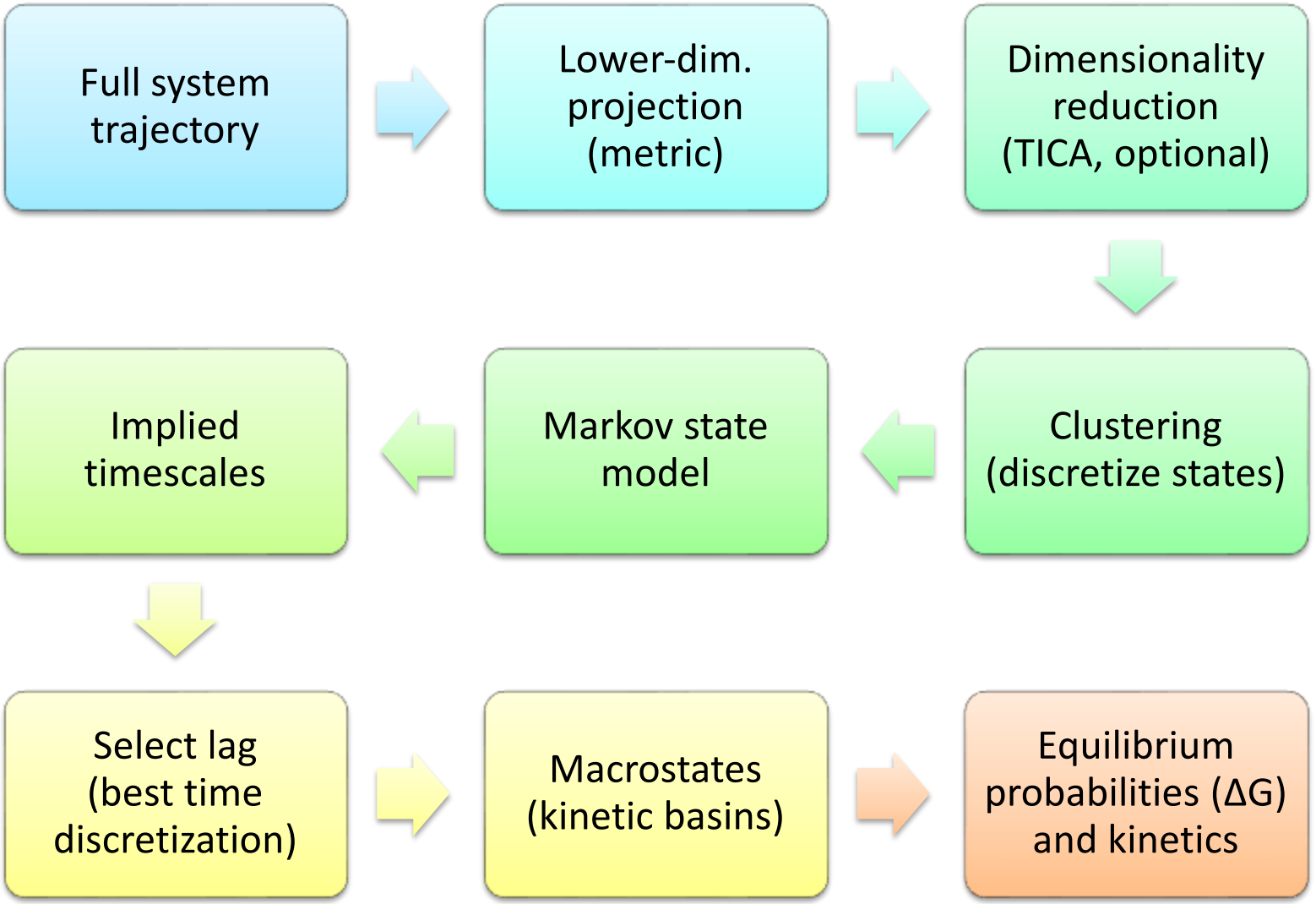
Hopefully better

* Better tests are available, e.g. Chapman-Kolmogorov

Back to molecules:

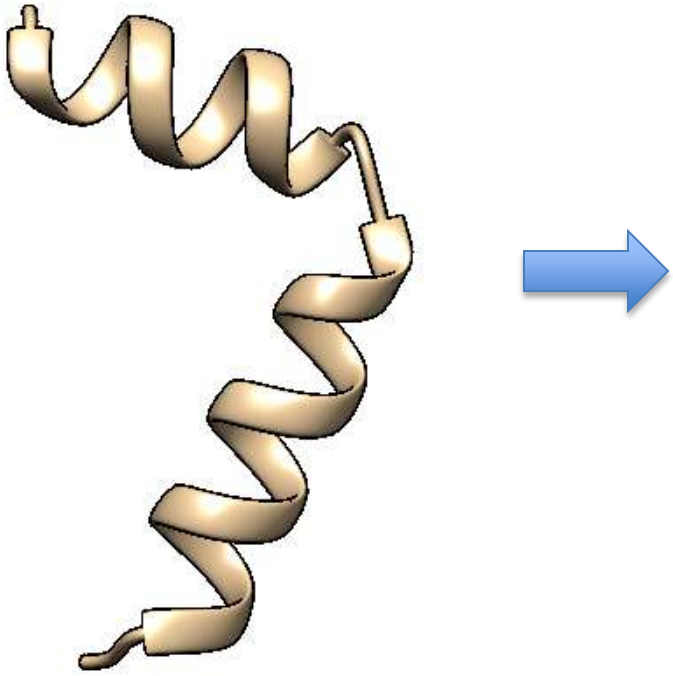
>> 1 dimension

MSM-based analysis overview



Metric projection

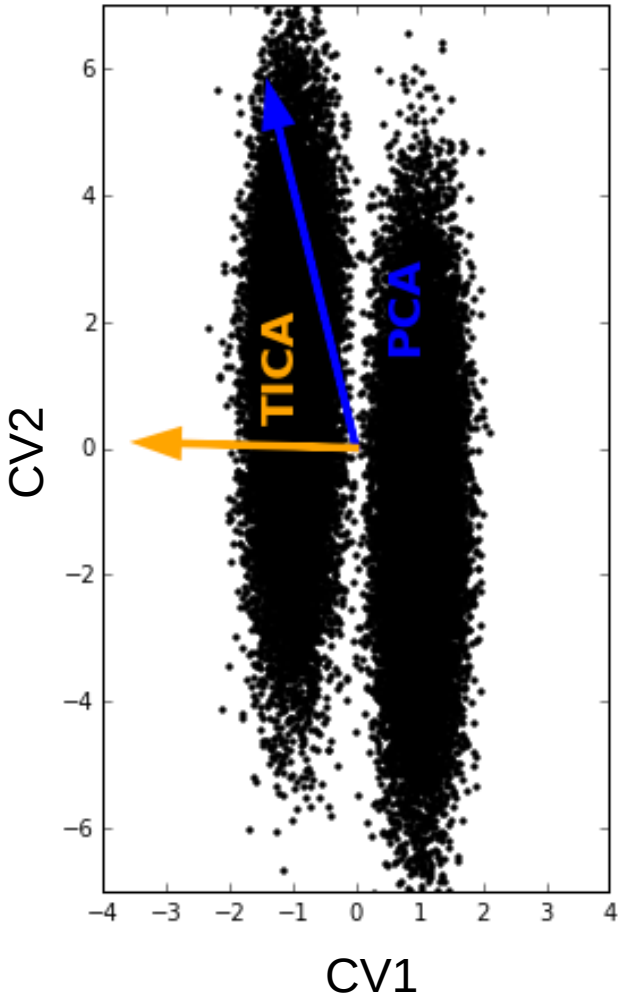
The first step is to project the system state in a lower-dimensional space ("metric"). Many choices are available, e.g.



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

Time-lagged independent component analysis

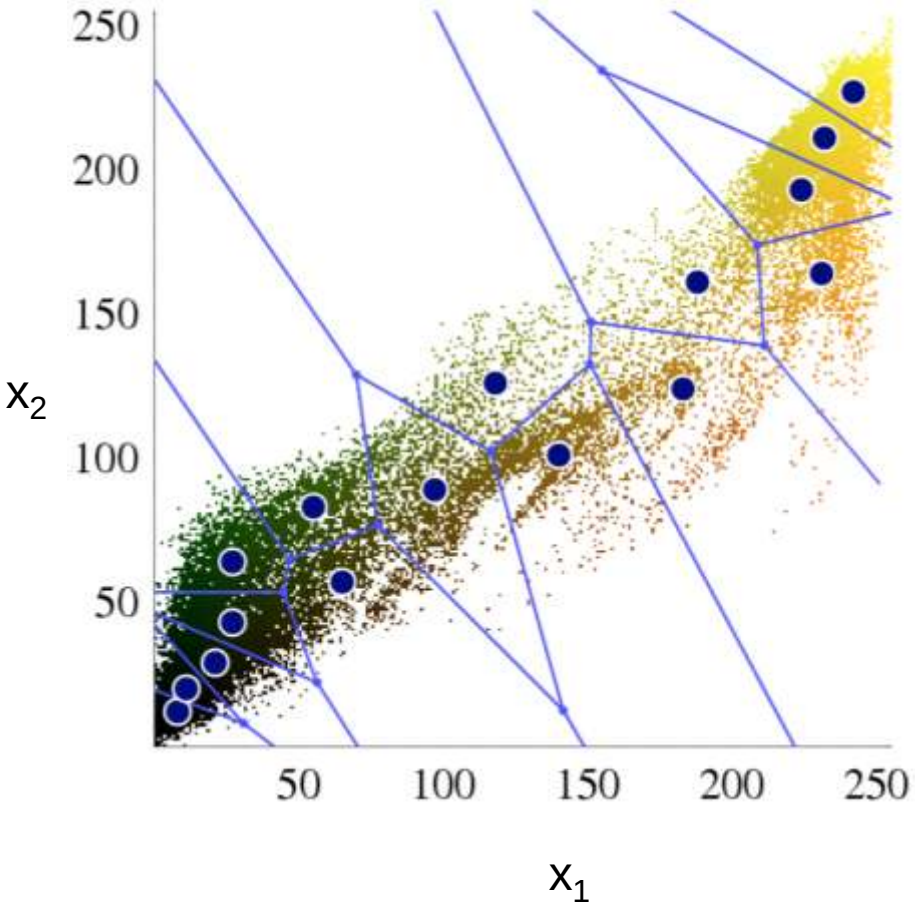
- The feature space dimensionality may still be too high
- Do a further low-dimensional projection on the "slow" degrees of freedom: TICA
- It is based on lagged autocorrelation
- Contrast with PCA, which fits the "most elongated" ellipsoid, ignoring time



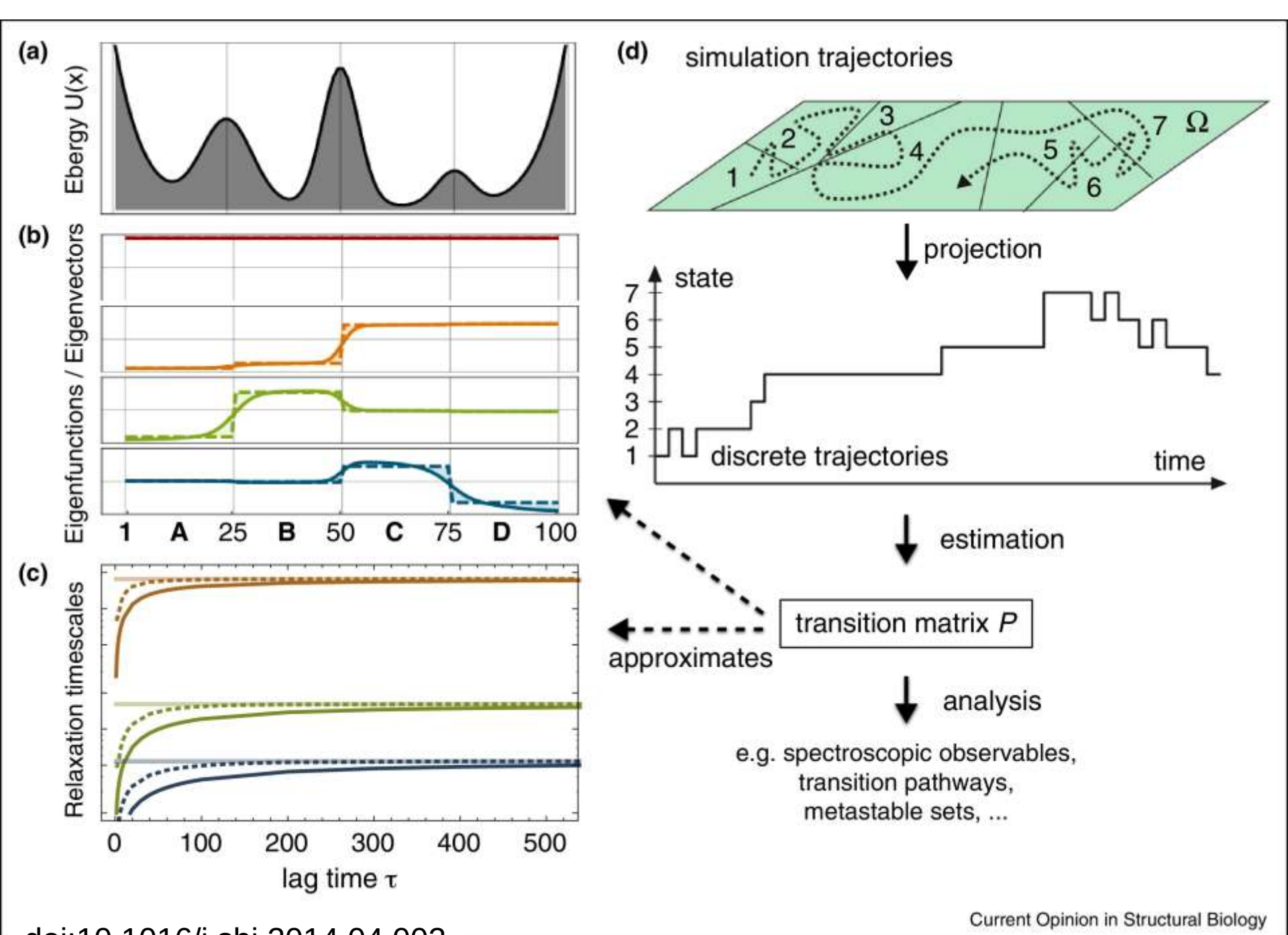
J. Chem. Phys. 139, 015102 (2013).

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

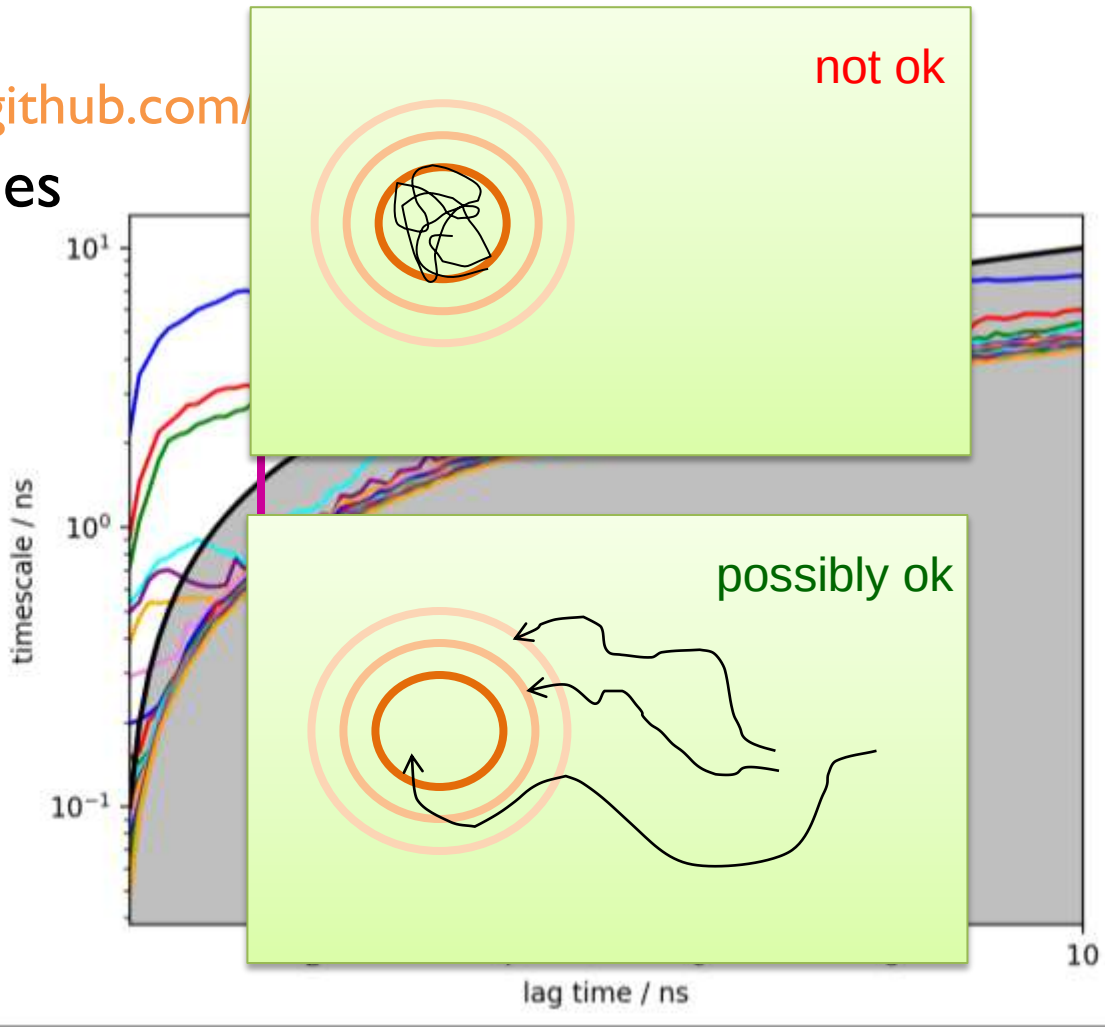


Current Opinion in Structural Biology

doi:10.1016/j.sbi.2014.04.002

Even when merging multiple trajectories, making efficient use of sampling does not dispense us from sampling the phase space

- Ace-Ala3-Nme
- Part of examples at: github.com/
- 6 Ramachandran angles
- The system is trapped
- How to «shoot» trajectories:
  - «bathtub» not ok
  - «shower» maybe
  - adaptive spawning
  - or your favourite string-like method



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

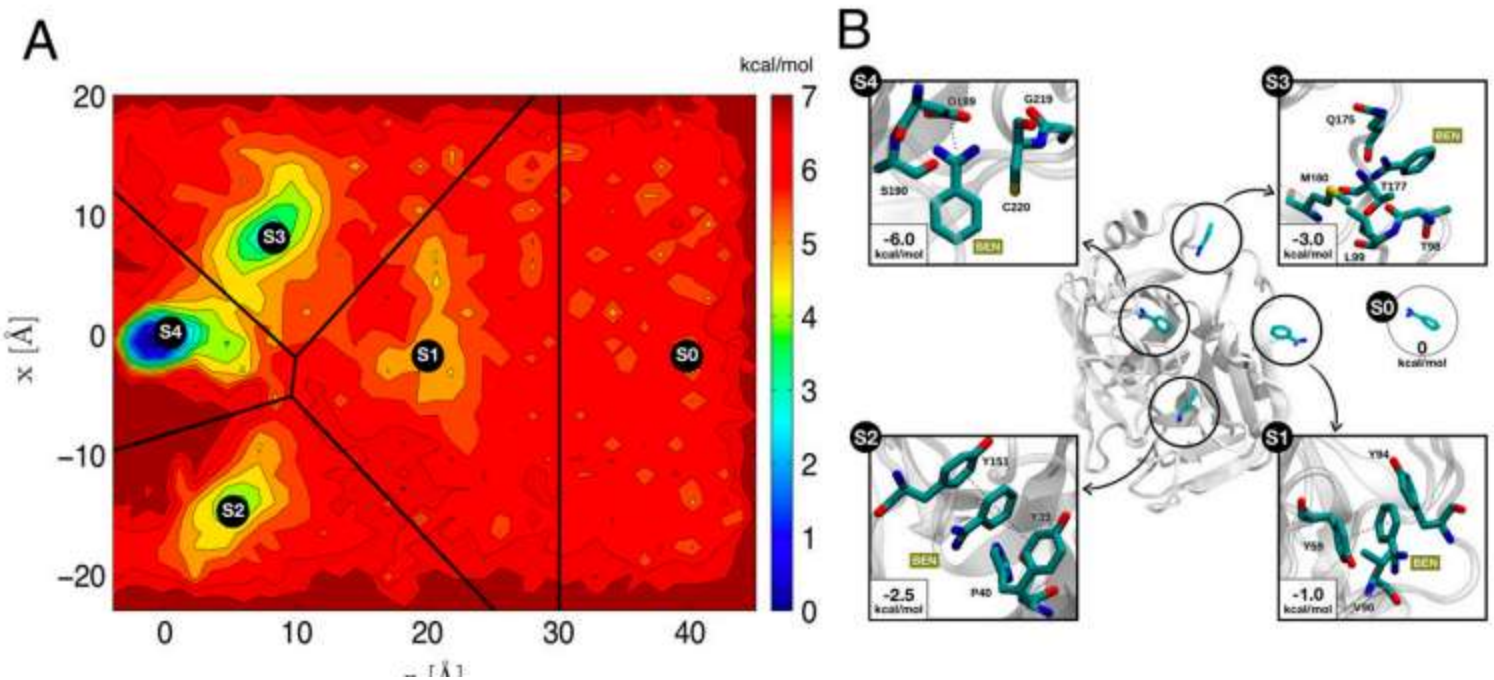
Edited by Arieh Warshel, University of Southern California, Los Angeles, CA, and approved May 11, 2011 (received for review March 4, 2011)

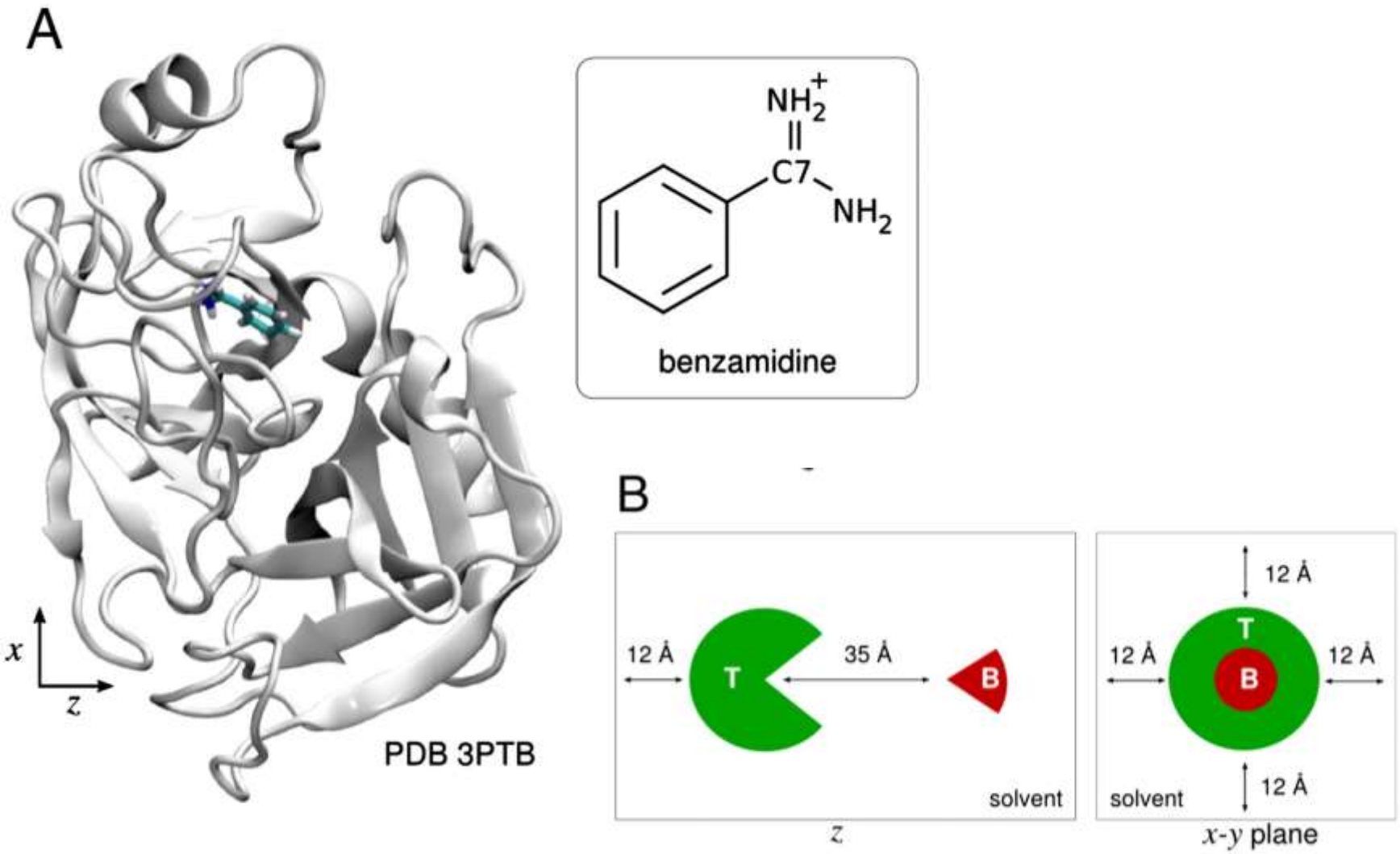
The understanding of protein–ligand binding is of critical importance for biomedical research, yet the process itself has been very difficult to study because of its intrinsically dynamic character. Here, we have been able to quantitatively reconstruct the complete binding process of the enzyme-inhibitor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

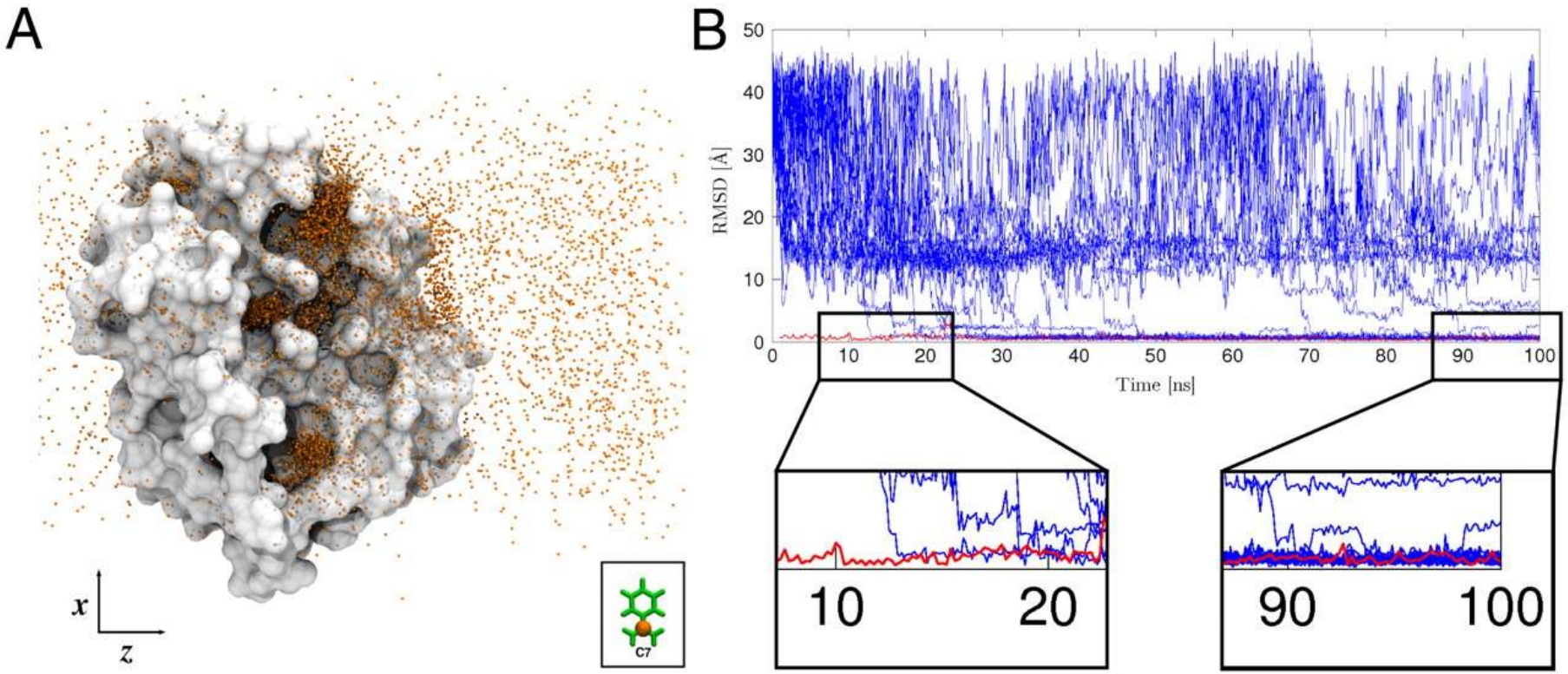
Free ligand binding has been used in the past to describe computational experiments in which, typically, a ligand is placed at a certain distance from the target protein and first by diffusion and

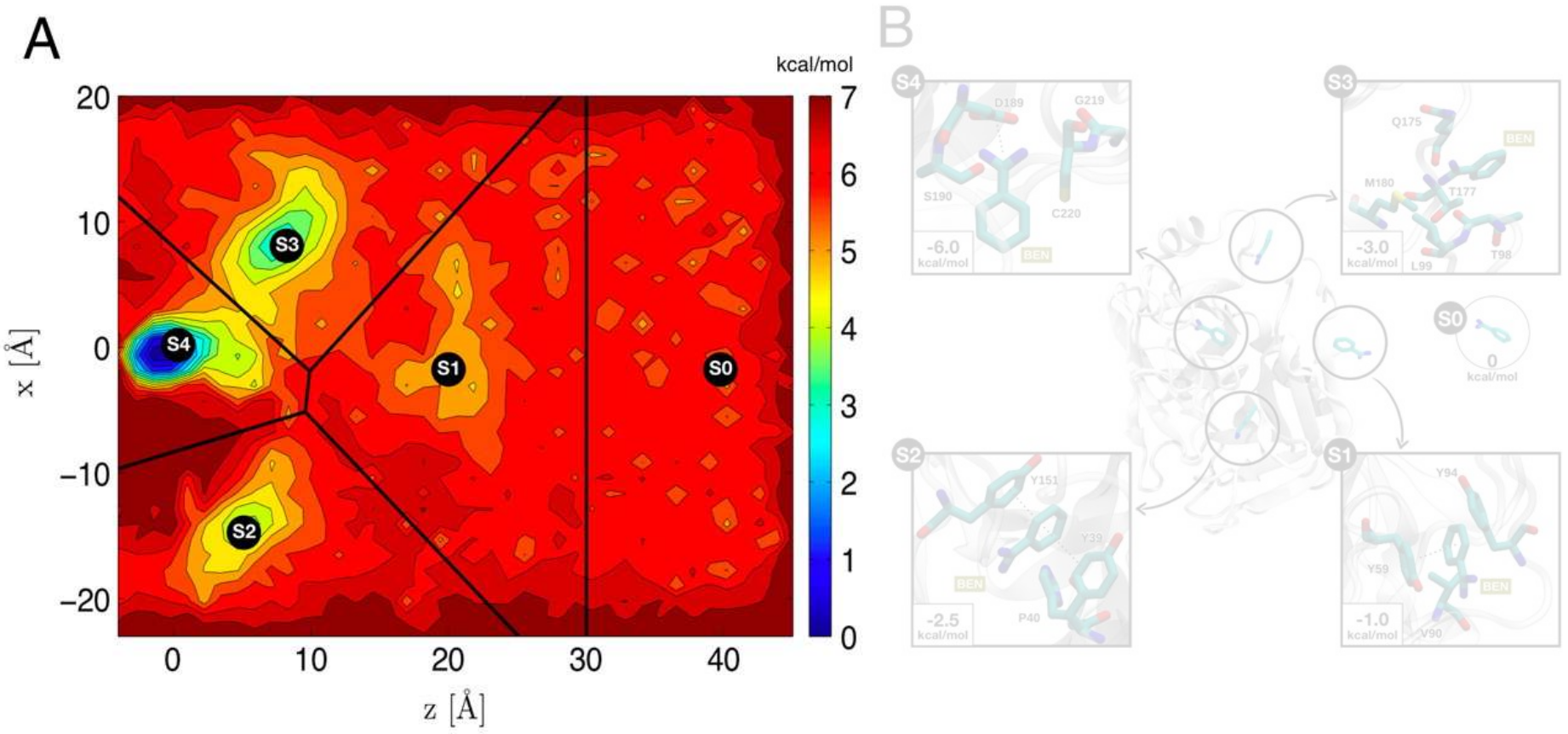
n the pro-





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



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

PNAS

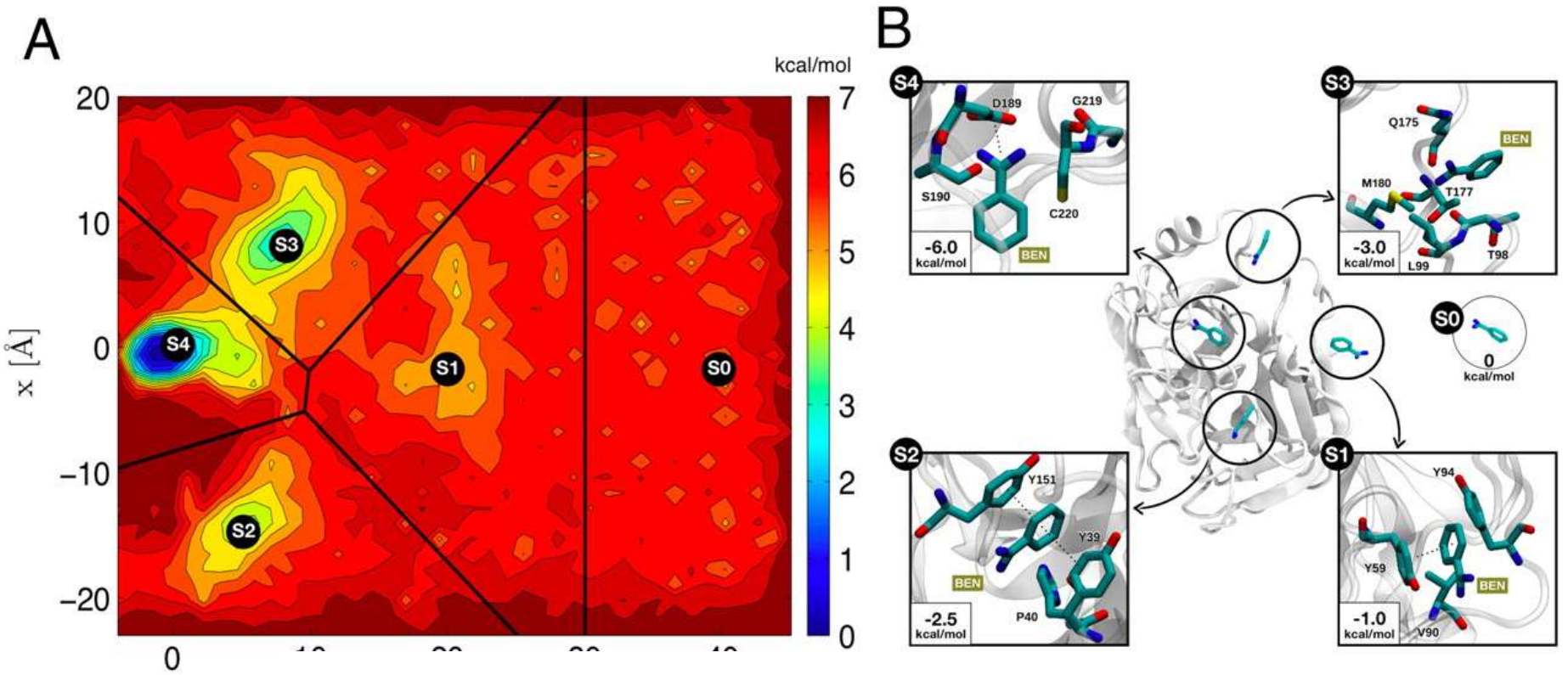


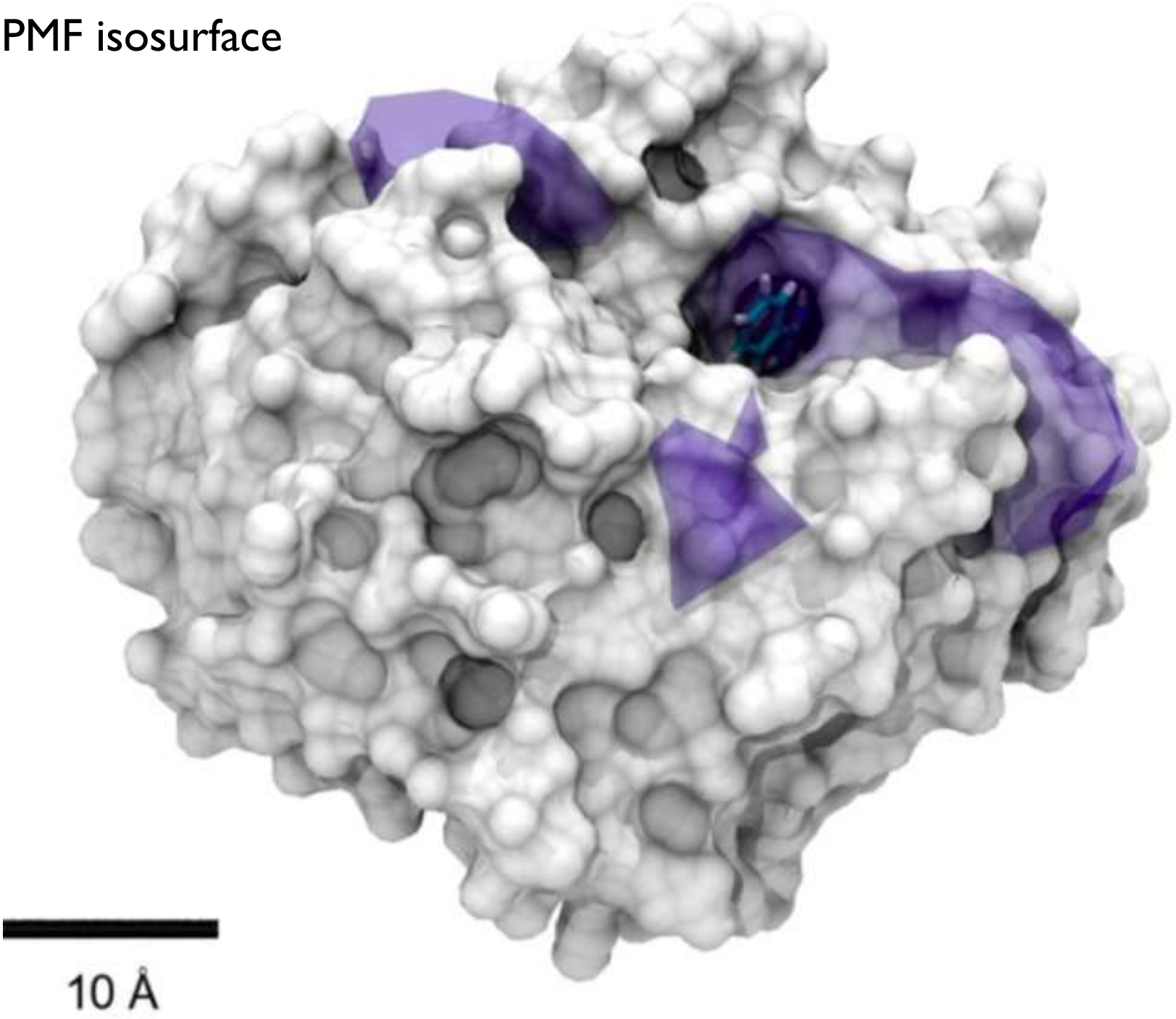
Table S1. Transition probabilities for the metastable states in the five-state coarse-grained model at a lag time of 50 ns

	S0	S1	S2	S3	S4
S0	0.069	0.090	0.130	0.305	0.406
S1	0.066	0.094	0.124	0.300	0.416
S2	0.059	0.091	0.186	0.313	0.352
S3	0.073	0.096	0.103	0.361	0.366
S4	0.021	0.032	0.043	0.107	0.797

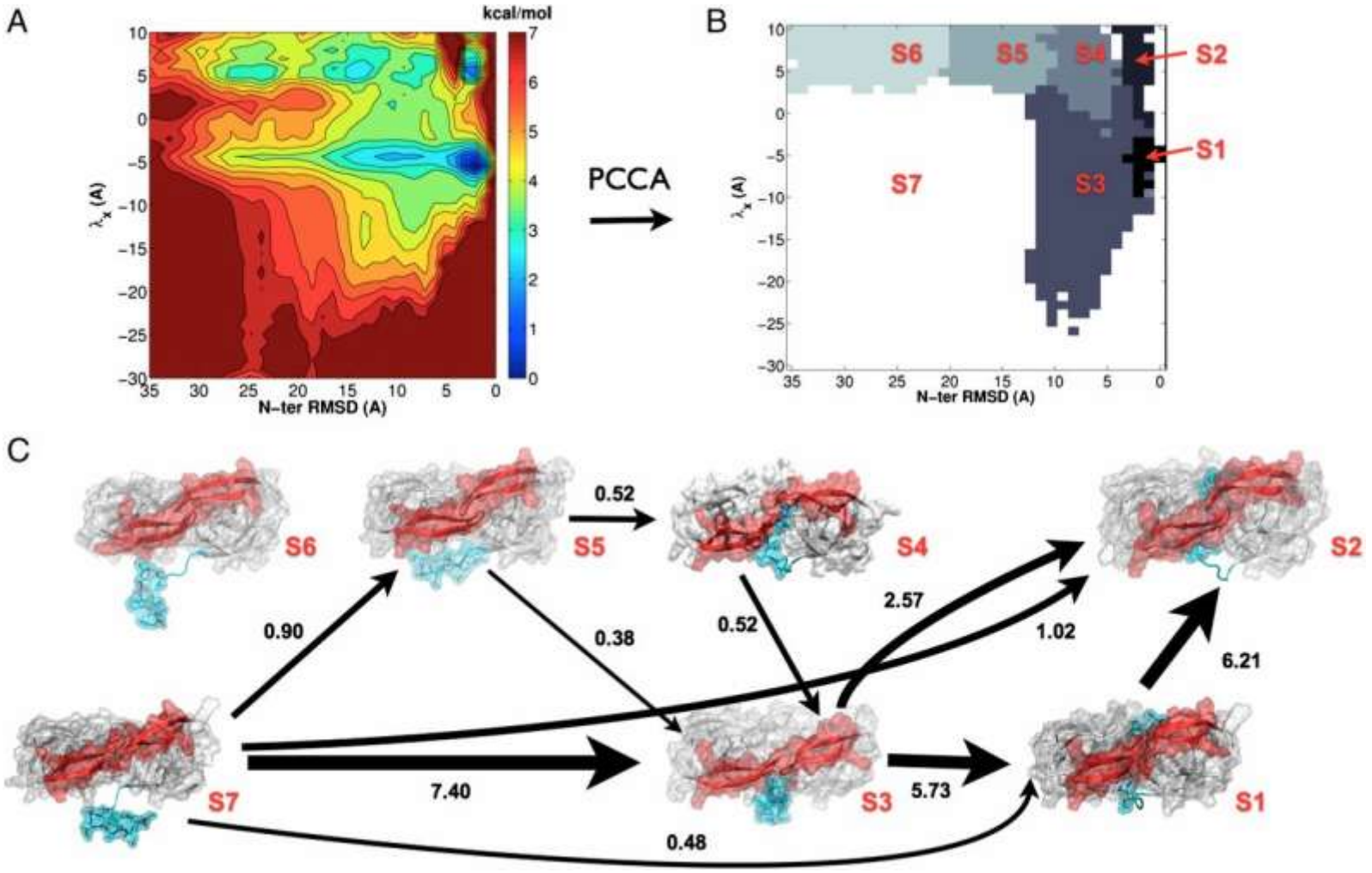
Identification of
identified from th
state S0 and the
from S3 from th
characterization
interactions with
of benzamidine (
between the Q17

states can be
en the unbound
tate S4 may be
(B) Structural
y π-π stacking
en NH2 groups
n-π interaction

3D PMF isosurface

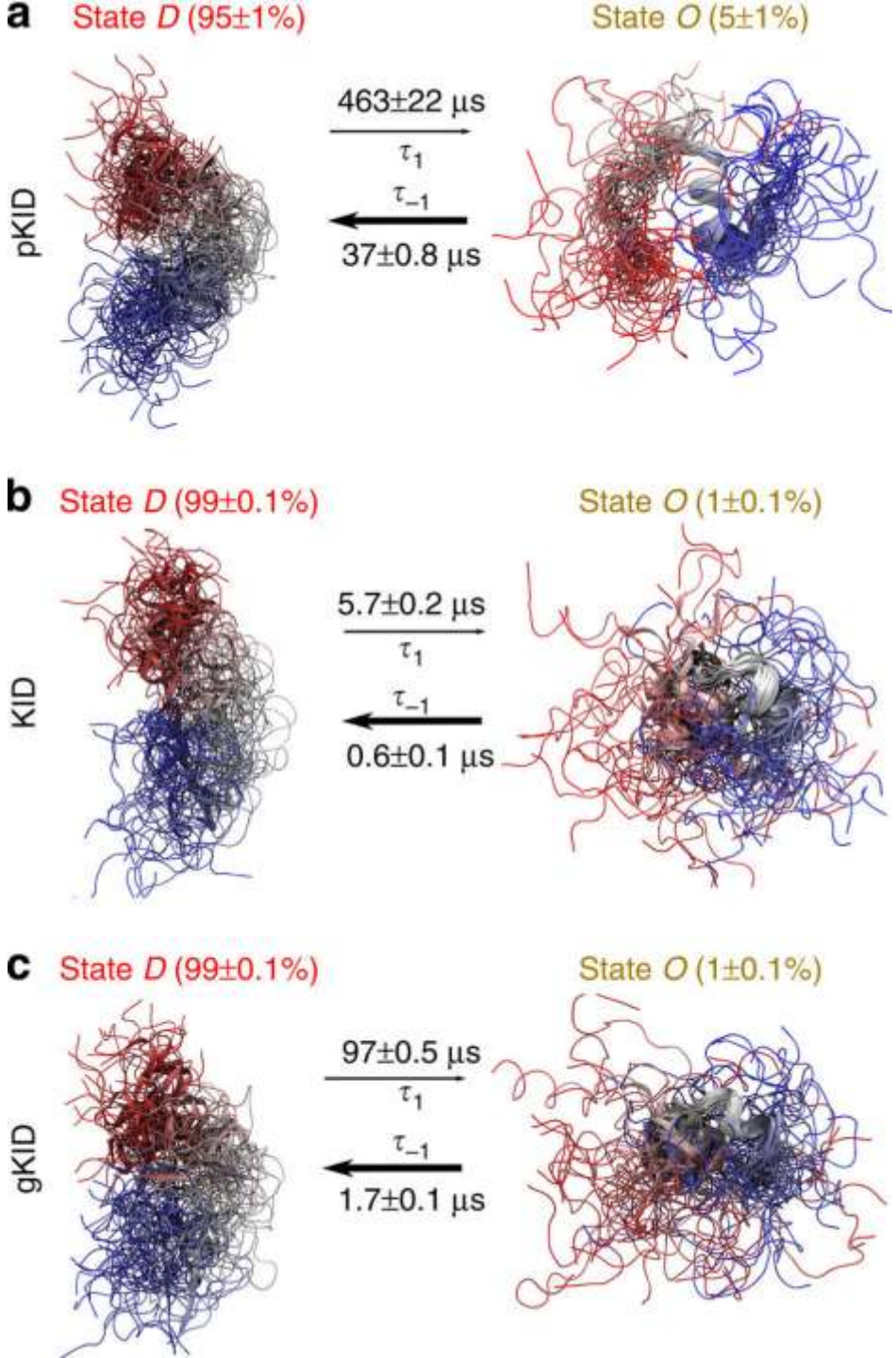


Kinetic characterization of the critical step in HIV-1 protease maturation



www.pnas.org/content/109/50/20449.full

Kinetic modulation of a disordered protein domain by phosphorylation



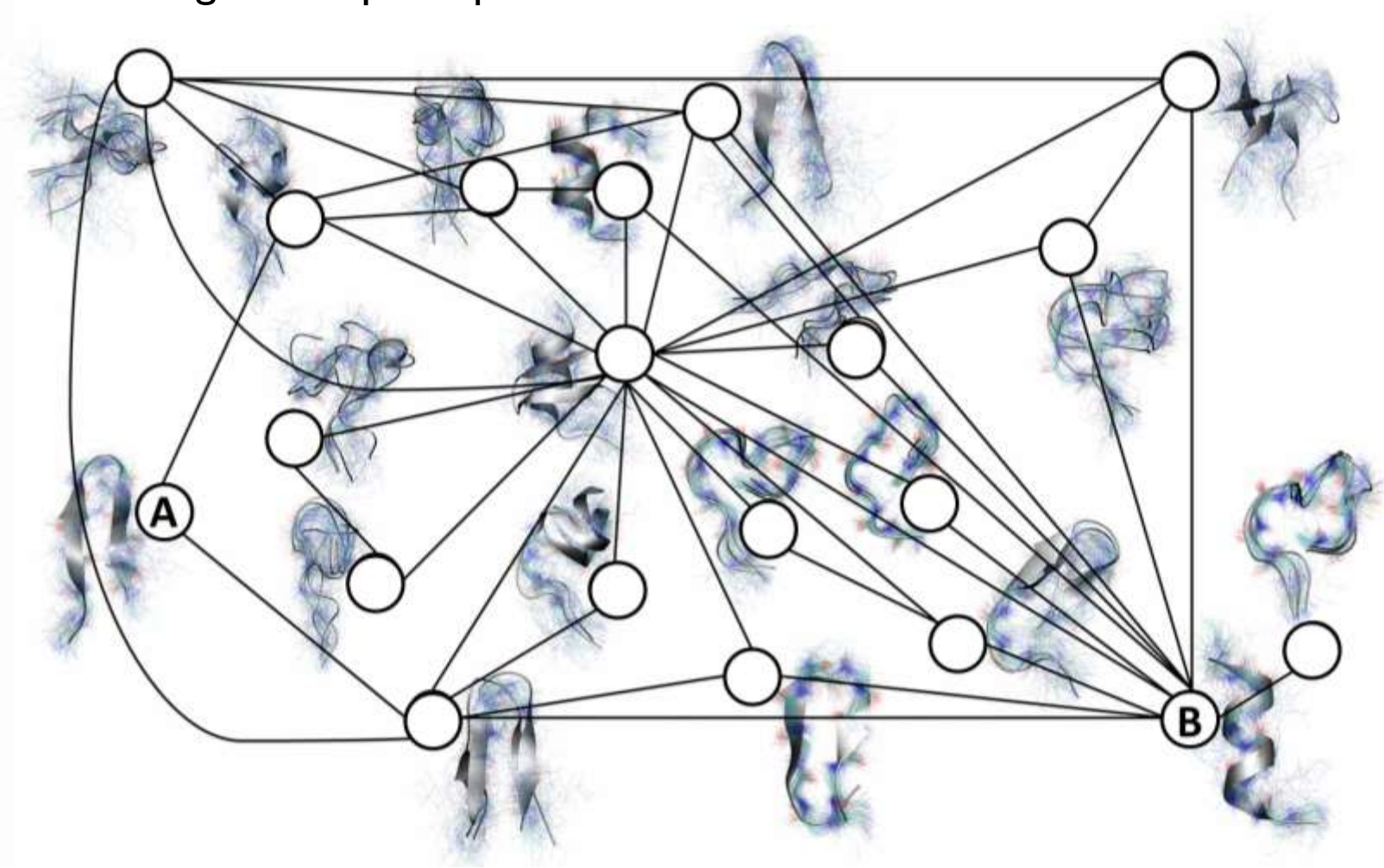
N. Stanley, S. Esteban and G. De Fabritiis, Nat. Commun. 5, 5272 (2014)

doi:10.1038/ncomms6272

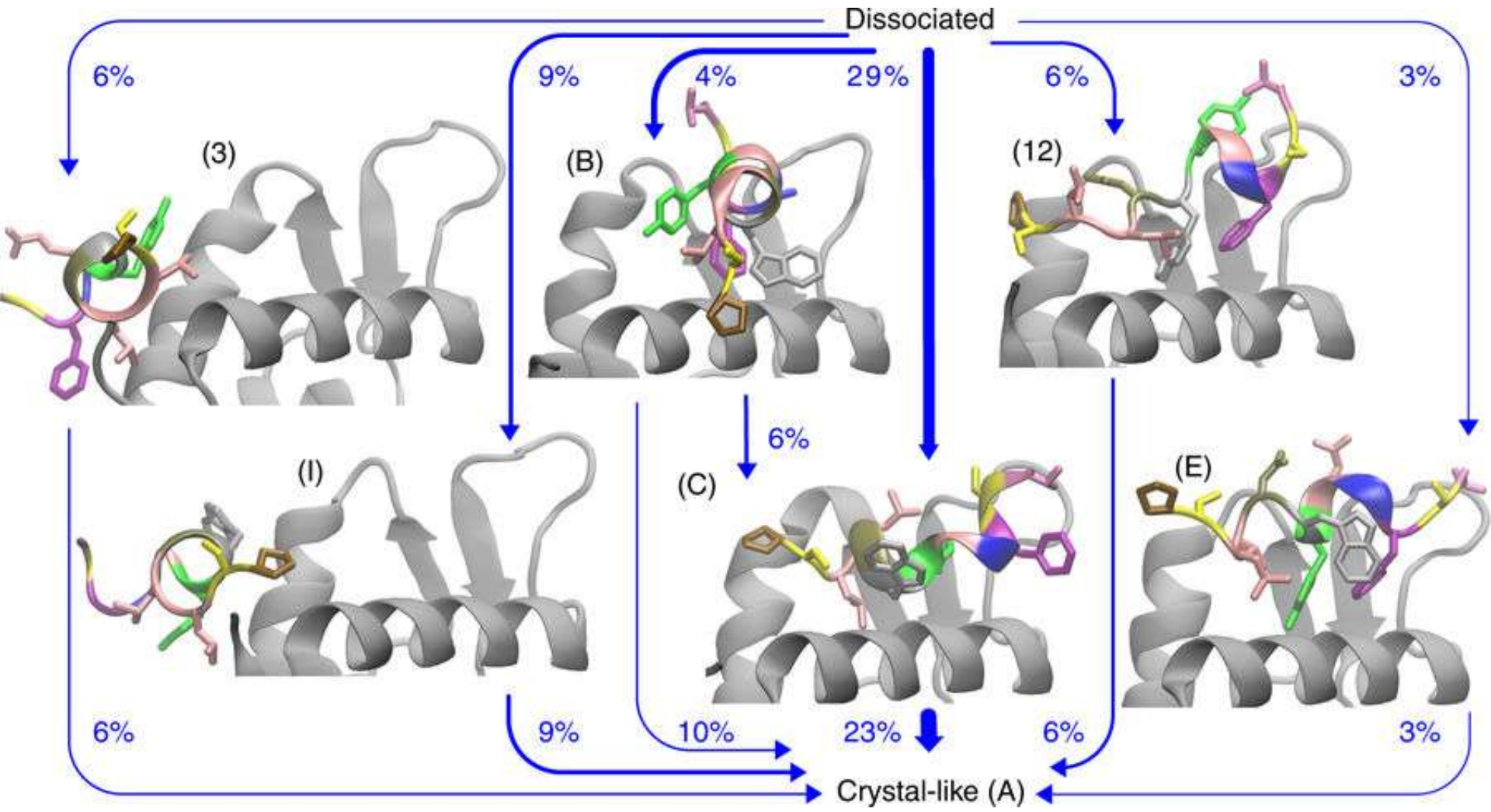
Folding kinetics

It is possible to model (small) protein folding processes at all-atom resolution.
However, the folding landscape is quite convoluted!

$k_{on} = ?$
$k_{off} = ?$



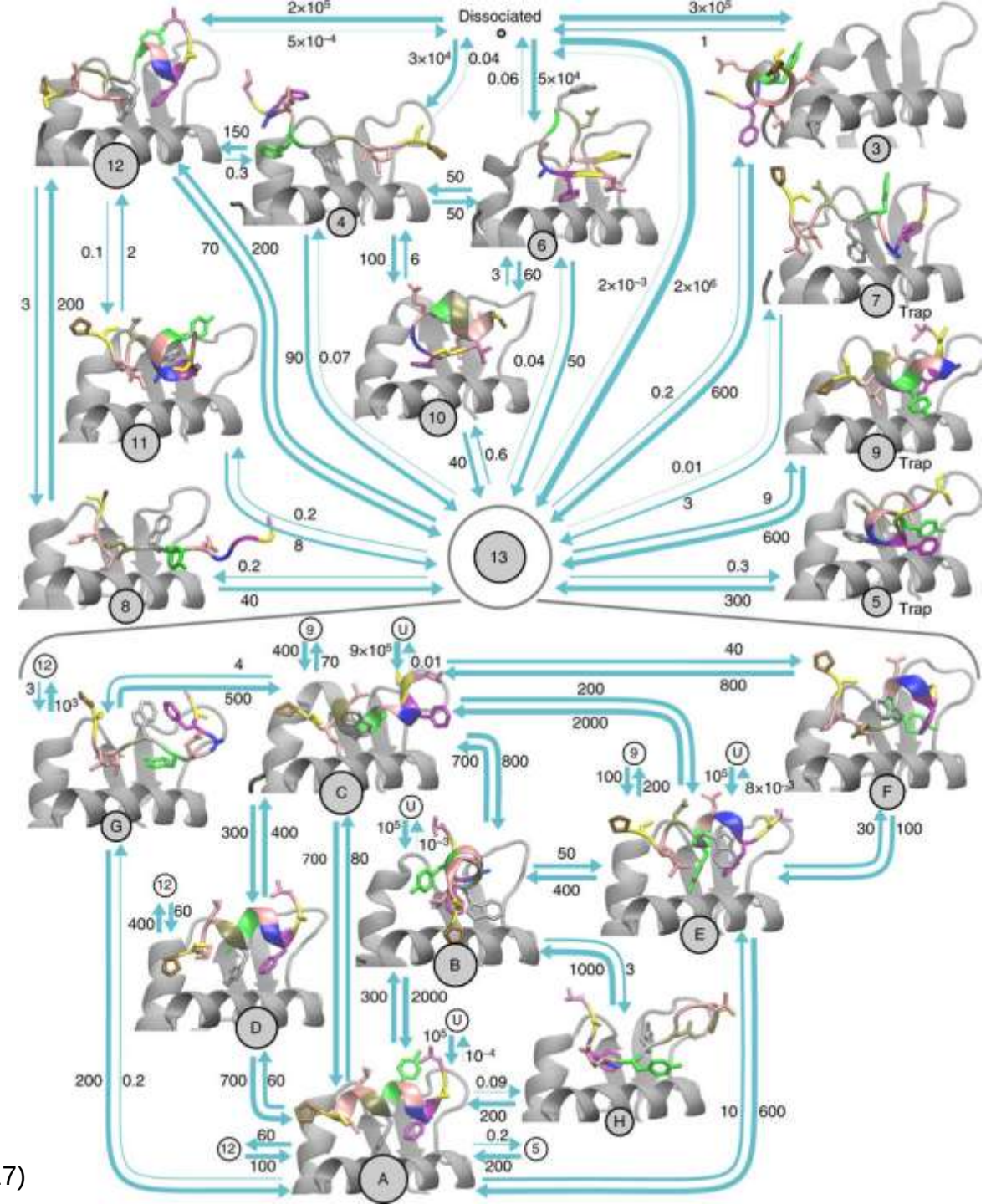
Banisch et al., ZIB-Report 15–19 (March 2015)



Binding mechanism comprised by the 60% most probable pathways. Structures of metastable (on-pathway) intermediates are shown, labels are as in Fig. 1. Arrows indicate the direction and relative magnitude of the reactive flux from the dissociated state to the crystal-like bound state. PMI residues that form PMI–Mdm2 contacts with at least a probability of 0.5 in a given macro-state are shown as sticks

Paul et al, Nat Comm, 1095 (2017)

Metastable states and transition rates for the binding of PMI to Mdm2. The PMI peptide is colored according to (). States are represented by discs with areas proportional to the natural logarithm of the equilibrium probability. Arrows indicate transitions with rate constants of at least 1 ms^{-1} in either direction. Numbers quantify transition rate constants in $ms^{-1} M^{-1}$ for association events and in ms^{-1} for all other transitions. The definition of the states is hierarchical: between top-level states 0 and 13, transitions happen on timescales of 10 µs or slower. States in the lower part of the figure are sub-states of top-level state 13. There, PMI transitions between different states in the main binding pocket of Mdm2 on timescales of microseconds or slower (only states with large probabilities are shown)



Paul et al, Nat Comm, 1095 (2017)

Conclusions and Resources

Conclusions

- MSM methods may be an attractive formalism for medium-sized problems
- Make efficient use of *unbiased* sampling
- Still require *huge* (but not unreachable) 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

Related topics

If you liked this, you may also like…

- Estimation of reversible transition matrices
- Hidden Markov Models doi:10.1063/1.4828816
- TICA
- Conformational fluxes
- Adaptive sampling
- Role of drug residence time in drug discovery

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
- Pande et al., 10.1016/j.ymeth.2010.06.002
  - Excellent overview (2)
- Paul et al., 10.1038/s41467-017-01166-6
  - 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