Markov-state modeling of biomolecular systems pt. 2

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

@giorginolab

Projects available!

Master in Bioinformatics for Health Sciences
Barcelona, 26 Apr 2019

Instructions

- Open github.com
- GIT clone (or unzip) **giorginolab/Markov-Tutorial-BCN-2019**
- This will copy the exercise files in your machine

Plan of the practice

1. We compute a MSM of a "toy model": trajectory in a 1-D potential (R language).
2. (Advanced) Move to a realistic trajectory set: prothrombin:inhibitor binding
   a. PLUMED projection
   b. build a simple MSM in R like before
3. (Self-study) Repeat the analysis with a ready-made library, HTMD

1. Toy model in 1D

- Start the R environment
- See the "*Markov From Scratch*" document
- Data file needed: data1.csv.gz
  - data10.csv.gz is also available for experimenting, see comments

```
$ zless data1.csv.gz
2.3000000000000000e+01
2.3000000000000000e+01
2.4000000000000000e+01
2.5000000000000000e+01
2.6000000000000000e+01
2.6000000000000000e+01
2.6000000000000000e+01
```

Let's now follow the R notebook called *1_MarkovFromScratch*

2. Prothrombin:ligand example

- This is a large, realistic simulation set
  - From htmd.org: Ligand Binding Analysis*
  - 3 GB (don't download today)
  - 852 trajectories (in 3 groups)
  - 852 x 20 ns ≈ 17 μs
- See the *MarkovOnLargeDataset* document

* http://pub.htmd.org/tutorials/ligand-binding-analysis/datasets.tar.gz

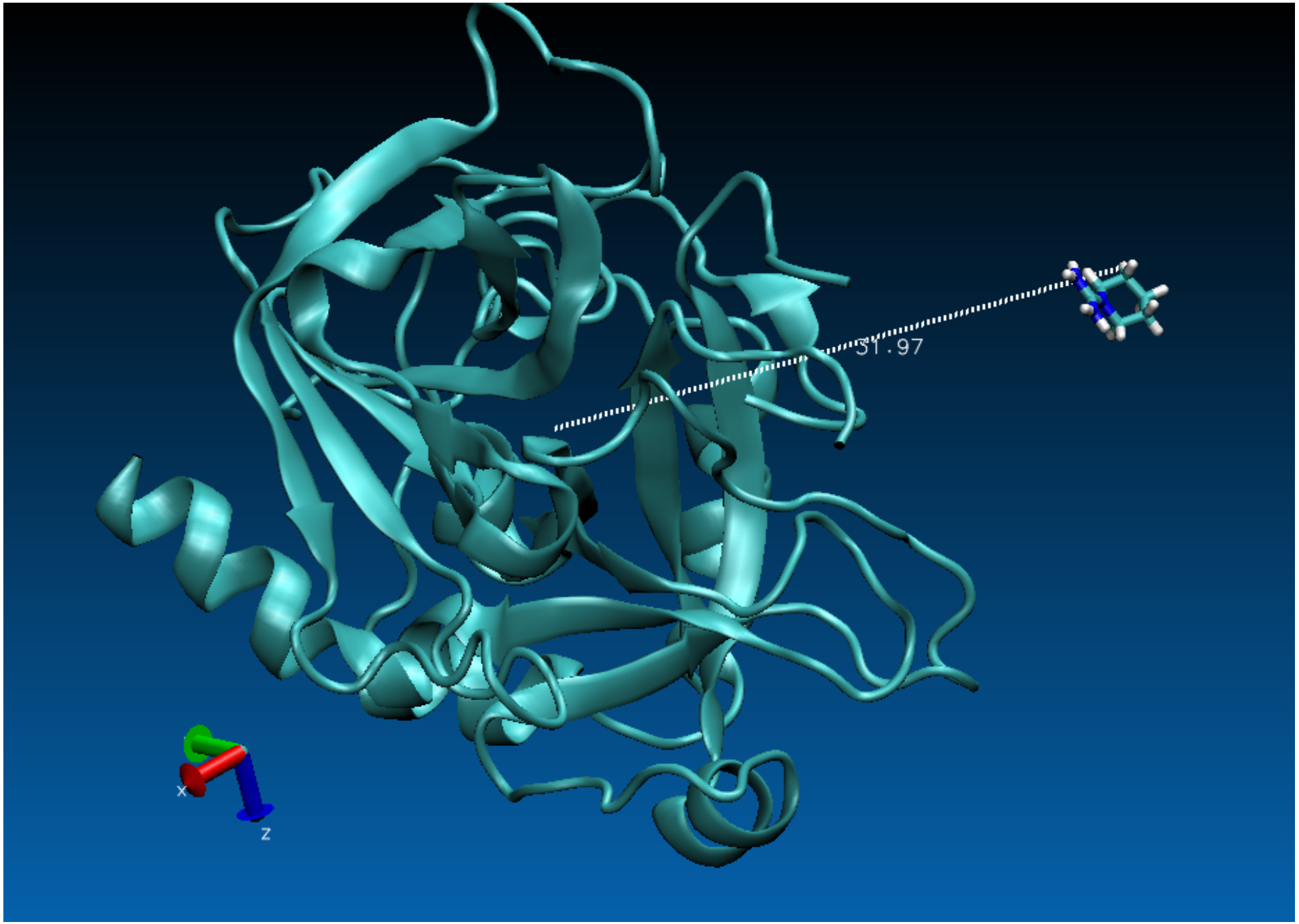
MarkovOnLargeDataset

- It is an R notebook similar to the 1-D example. However…
- The original space is high-dimensional. So
  1. Project it with a PLUMED script: Protein-Ligand vector after alignment
  2. Convert the 3D projection to discrete states with a *k-means* clustering algorithm

Using PLUMED for projections

- We need to extract from *each frame of each trajectory* a limited number of values (projection, or metric)
- Must be orientation-independent (align)
- PLUMED* may be used for the task
  - protein-ligand-vector.plumed
  - project-with-plumed.sh
- Results: metric.dat

* http://www.plumed.org


31.97
x
z

protein-ligand-vector.plumed

```
UNITS LENGTH=A

FIT_TO_TEMPLATE STRIDE=1 REFERENCE=reference.pdb TYPE=OPTIMAL

# These are the serial numbers of protein CA's
prot: CENTER ...
ATOMS={   5,   20,   30,   42,   52,   59,   78,  110,  116,  135,  155,
        170,  192,  214,  225,  244,  259,  271,  293,  307,  322,  346,
                                      . . .
       4229, 4253, 4272, 4294, 4316, 4340, 4359, 4376, 4398, 4414, 4433,
       4445, 4462
       }
...

# This is "resname MOL and noh" (inhibitors's heavy atoms)
ligand: CENTER ATOMS=4481,4484,4487,4490,4493,4496,4497,4498,4501

pl: DISTANCE ATOMS=prot,ligand COMPONENTS

PRINT ARG=pl.*
```

project-with-plumed.sh

Computes the metric over all the trajectories

(You don't actually have to run this. The results are already in **metrics.dat**)

```
#!/bin/bash

# This file computes the PLUMED metric indicated in $script
# on all of the trajectory files in $indir. The results
# are printed in stdout.

# You only need to run this file if you change the definition
# of the metric. The current results are in metric.dat

script=protein_ligand_vector.plumed
indir=/mnt/scratch/shared/markov/binding/1/filtered

# Loop over all files ending by .xtc
for tj in `find $indir -name \*.xtc -and -not -name .\*`; do
  # Output file name
  outname=`basename $tj`
  echo Running plumed on $tj >&2
  plumed driver --plumed $script \
    --pdb $indir/filtered.pdb \
    --mf_xtc $tj | \
       egrep "^ " | sed "s+^+$outname+"
done
```

```
project-with-plumed.sh > newmetric.dat
```

metric.dat

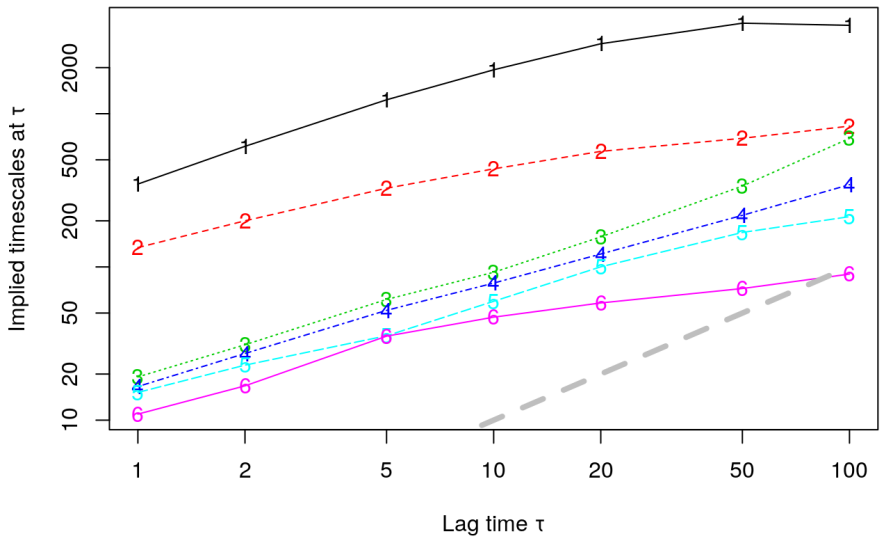
Results of the projection are in metric.dat.
This will be read in by R.

```
(htmd) tonigiorgino@monaco:~/practice$ head metrics.dat
e36s5_e27s3f37-....xtc 0.000000 -0.224091 -11.886612 1.973588
e36s5_e27s3f37-....xtc 1.000000 -0.413809 -12.038901 1.644167
e36s5_e27s3f37-....xtc 2.000000 -0.532973 -12.323975 2.144003
e36s5_e27s3f37-....xtc 3.000000 -0.300480 -11.363546 1.903810
e36s5_e27s3f37-....xtc 4.000000 -0.433555 -11.792027 2.057442
. . .
e24s3_e21s4f177-....xtc 196.000000 8.976495 -10.536748 2.060360
e24s3_e21s4f177-....xtc 197.000000 9.262883 -11.399111 2.169123
e24s3_e21s4f177-....xtc 198.000000 8.517335 -12.355884 2.464811
e24s3_e21s4f177-....xtc 199.000000 8.496182 -12.325769 1.489607
```

Let's now follow the R notebook called *3_MarkovOnLargeDataset*

Possible homework (I)

- We are starting to see convergence, but not quite. Use the PLUMED syntax to find better metrics that improve Markovianity.
- Explore the most stable states
- Explore major relaxation modes



Possible homework (II)

- Try the other split of the dataset. (1 vs 2 vs 3). Are they independent?
- Use all of them together.
- Try the other systems in, like `villin` and `cxcl12`
  1. Familiarize with the PDB
  2. Devise a reasonable metric

MD analysis libraries

Where to go from here (III)

- R is fine, but *specialized* environments for MD analysis are more productive
  - HTMD°, MdAnalysis, MdTraj, Bio3D, … *
  - Most are Python based
- ° It does:
  - Molecule/trajectory visualization
  - Markov Modeling (etc)

*

Analysis libraries for molecular trajectories: a cross-language synopsis

Toni Giorgino

October 1, 2018

To appear in:
Biomolecular Simulations: Methods and Protocols
Edited by M. Bonomi and C. Camilloni

Corresponding author's address/affiliation:
Biophysics Institute, National Research Council of Italy
Department of Biosciences, University of Milan
Via G. Celoria 26, I-20133, Milan, Italy

Running head:
MD analysis libraries: a synopsis

Summary

Analyzing the results of molecular dynamics (MD)-based simulations usually entails extensive manipulations of file formats encoding both the topology (e.g. the chemical connectivity) and configurations (the trajectory) of the simulated system. This chapter reviews a number of software libraries developed to facilitate interactive and batch analysis of MD results with scripts written in high-level, interpreted languages. It provides a beginners' introduction to MD analysis presenting a side-by-side comparison of major scripting languages used in MD, and show how to perform common analysis tasks within the VMD, Bio3D, MDTraj, MDAnalysis and HTMD environments.

1 Introduction

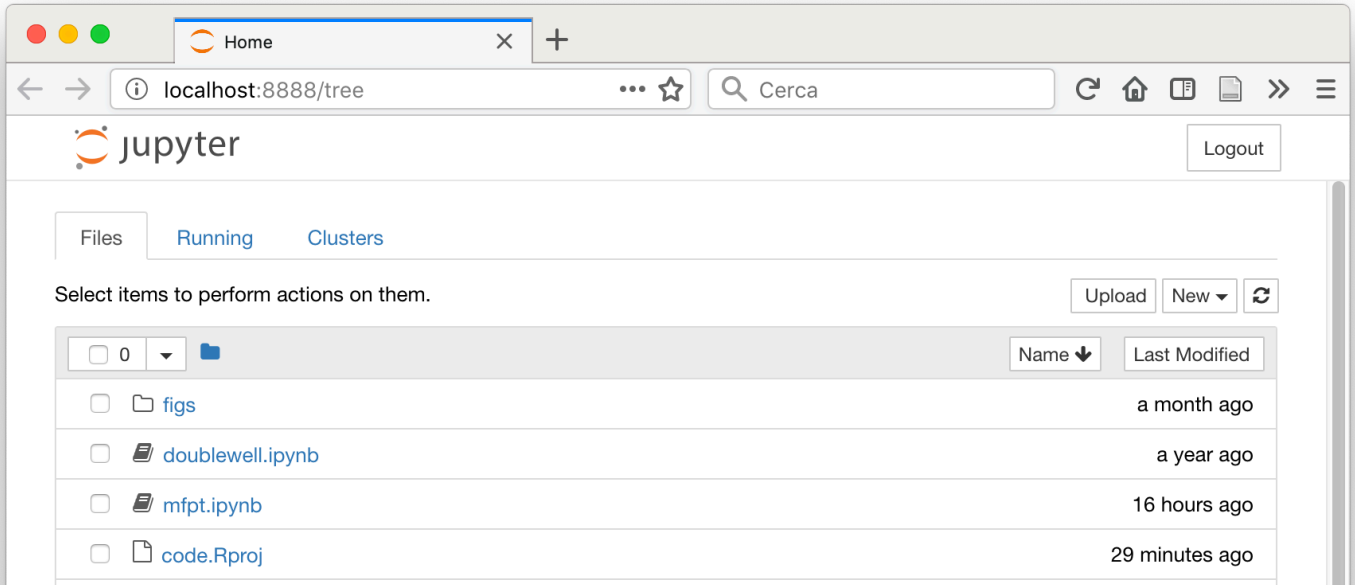
The backbone of molecular dynamics (MD) based methods is to integrate the equations of motion of a system with a given Hamiltonian. The integration is performed by an MD engine with a finite time-step, sufficiently fine to capture

1

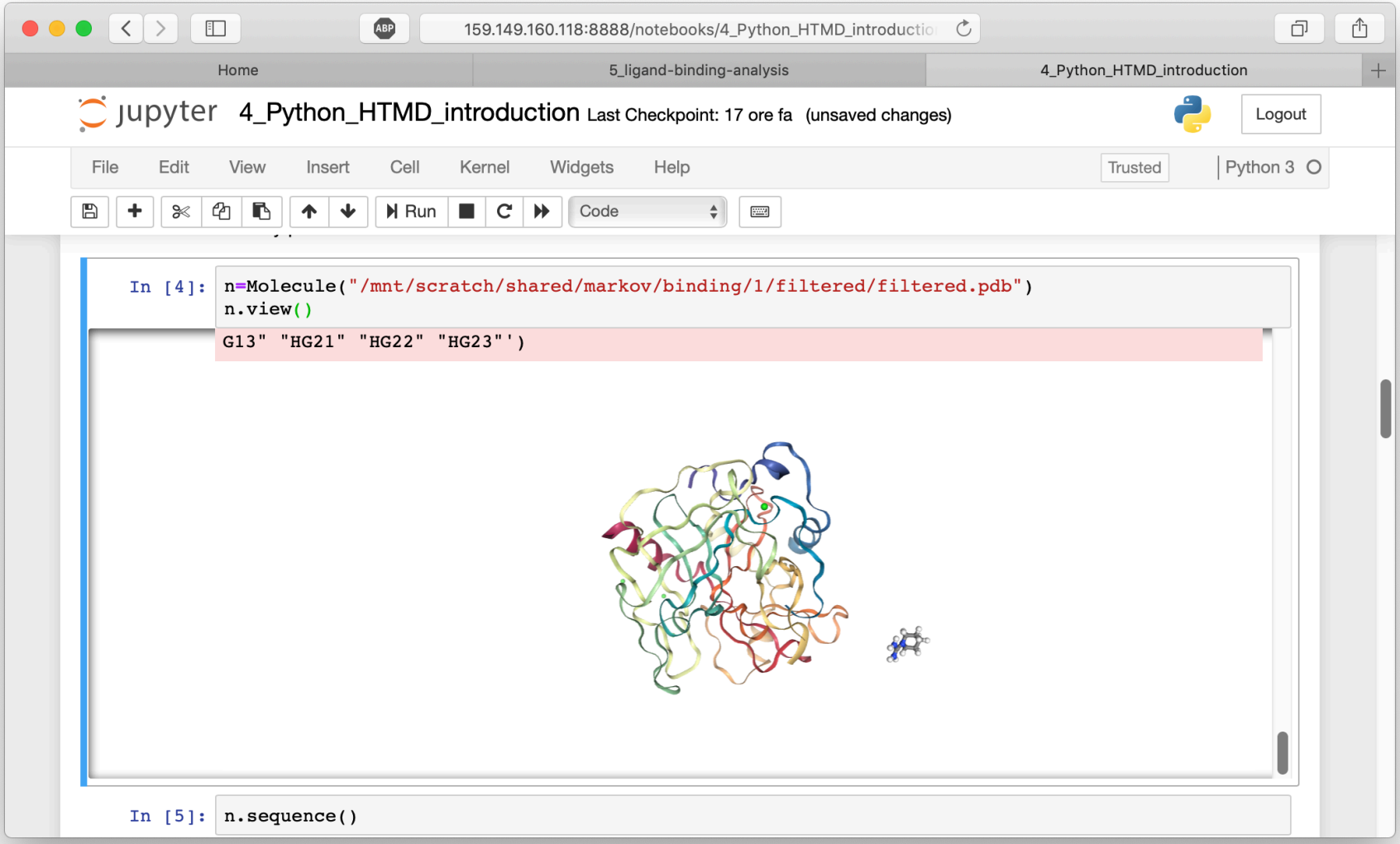
* GitHub: giorginolab/analysis_libraries_chapter

Python Notebooks

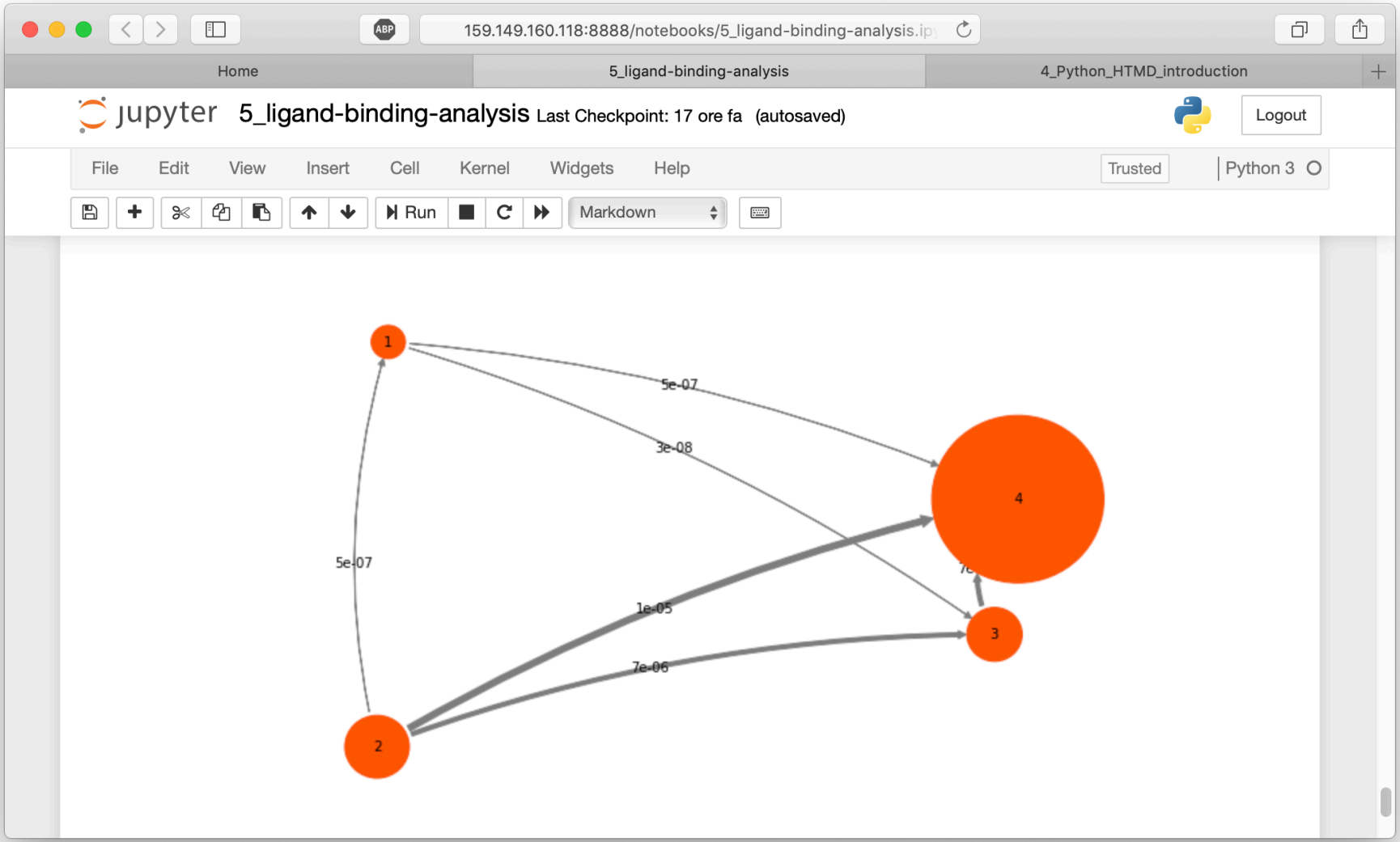
- Install the conda environment
- Start with "`jupyter notebook`"
- You will see an interactive Python (*jupyter*) Notebook.



Python Notebooks

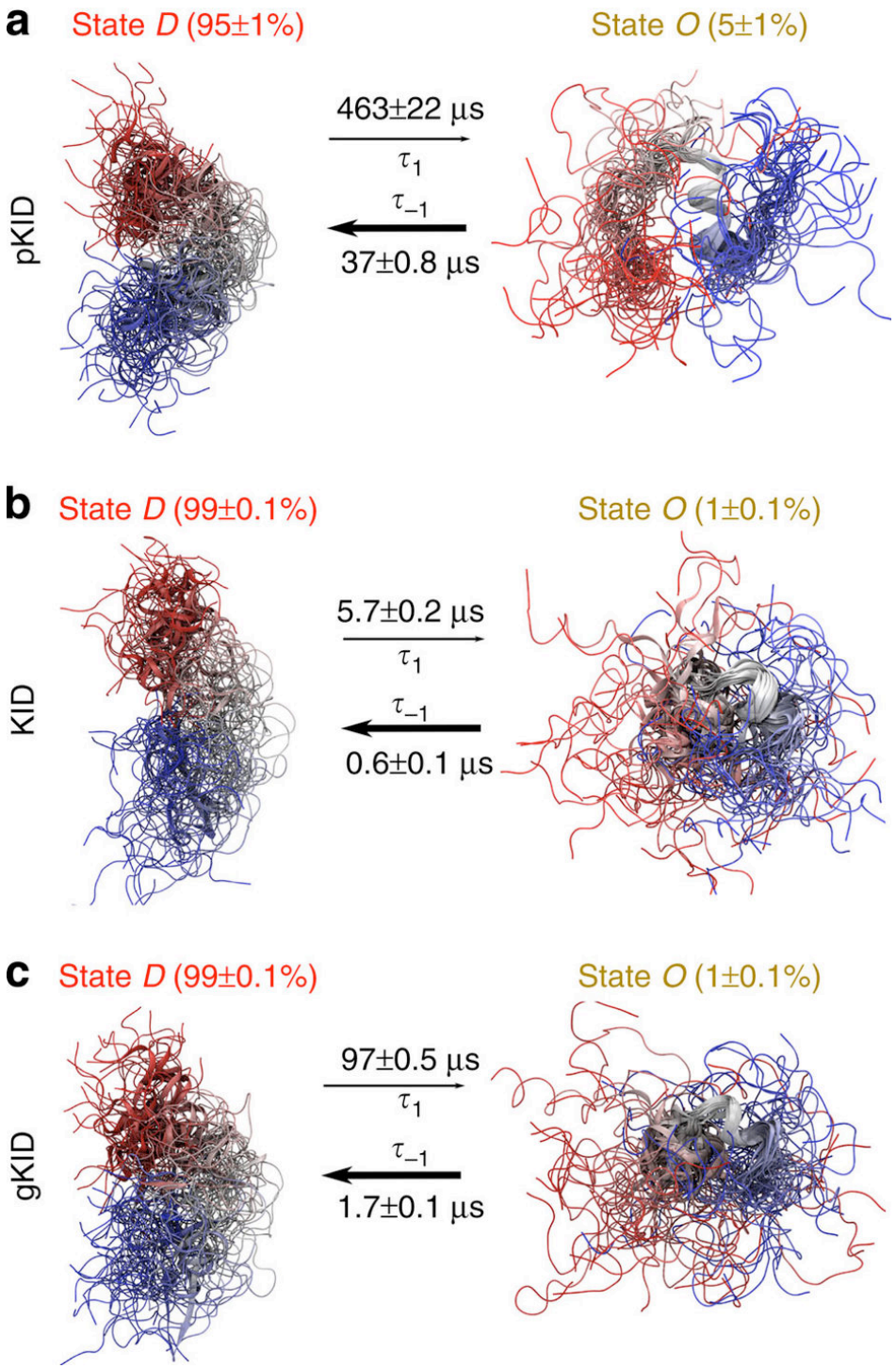


Python Notebooks (MSM)



Examples from the literature

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

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

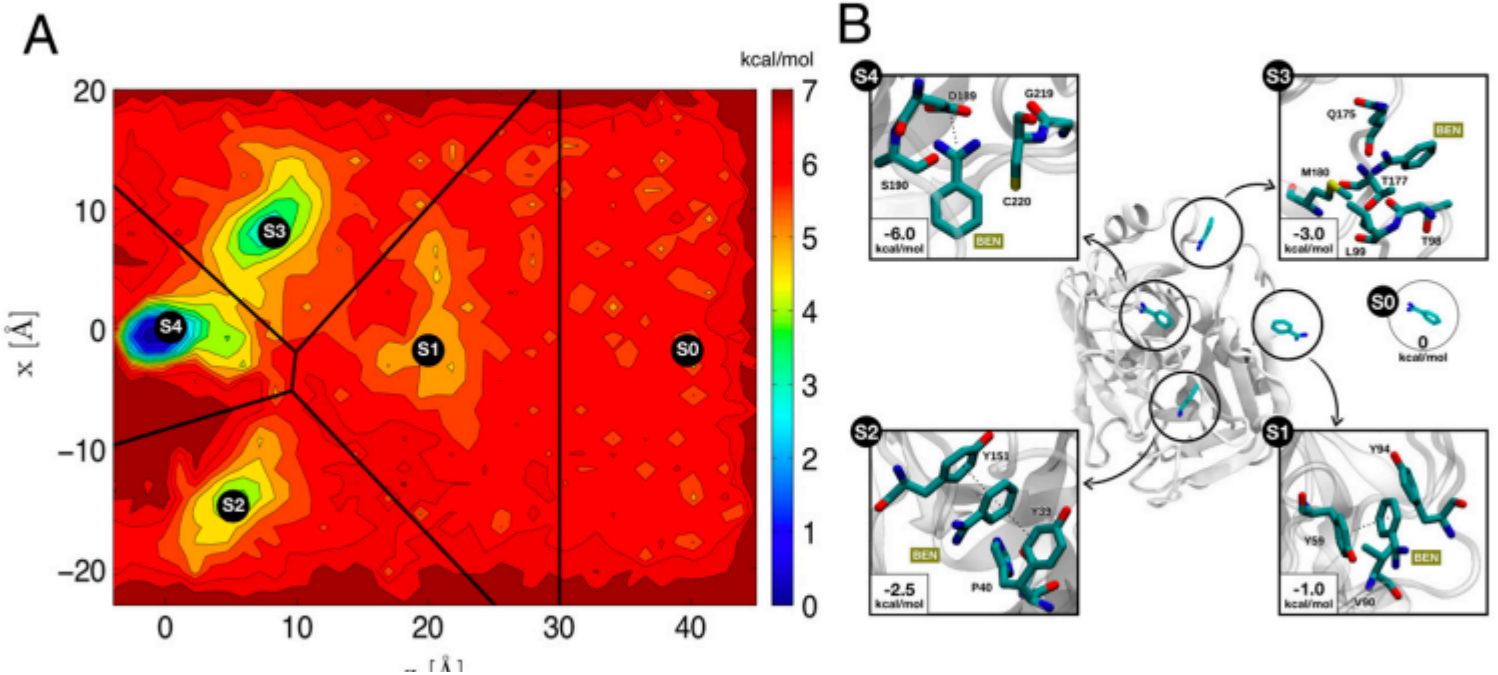
Edited by Arieh Warshel, University of Southern California, Los Angeles, CA, and approved May 11, 2011 (received for review March 4, 2011)

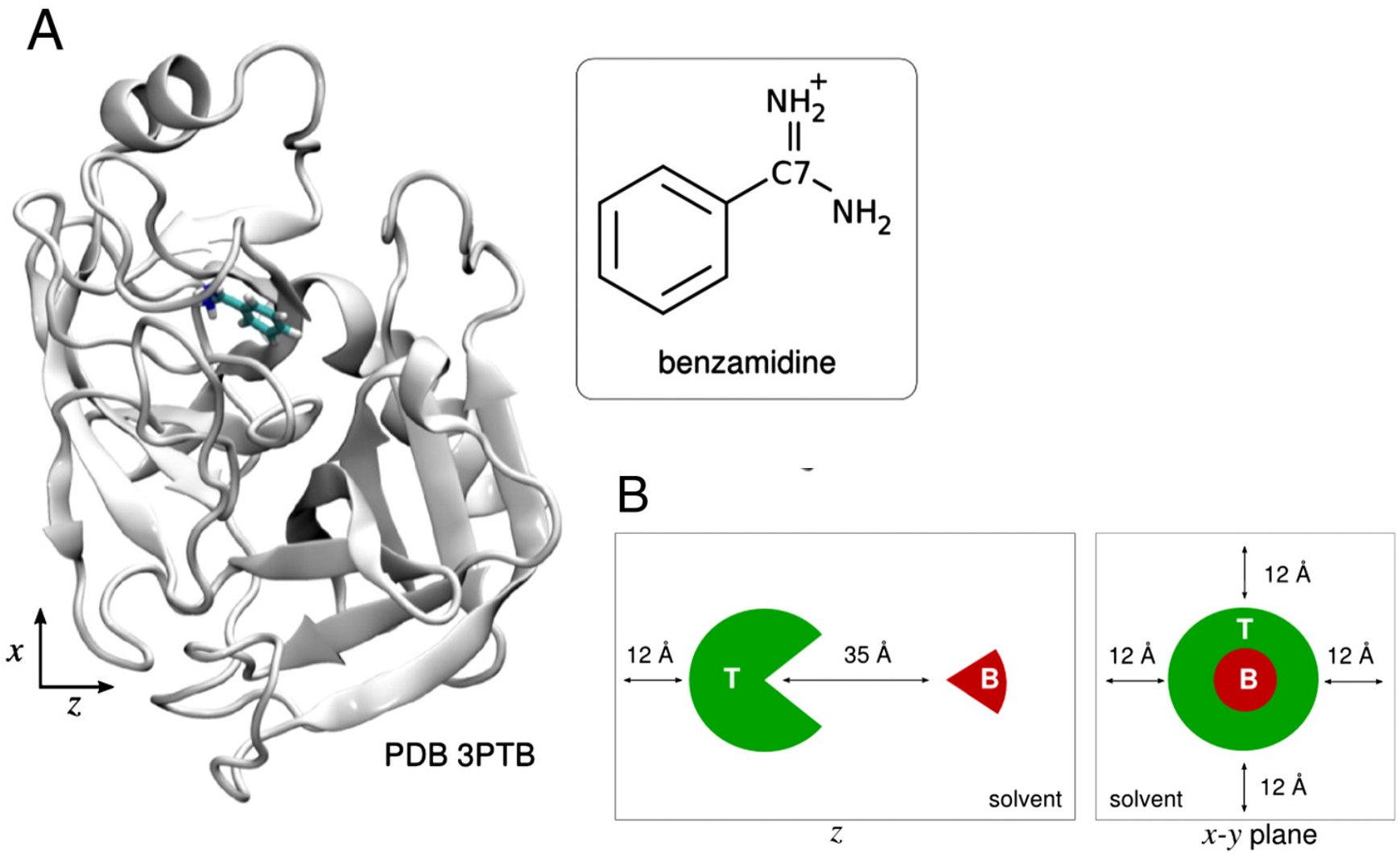
The understanding of protein–ligand binding is of critical importance for biomedical research, yet the process itself has been very difficult to study because of its intrinsically dynamic character. Here, we have been able to quantitatively reconstruct the complete binding process of the enzyme-inhibitor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

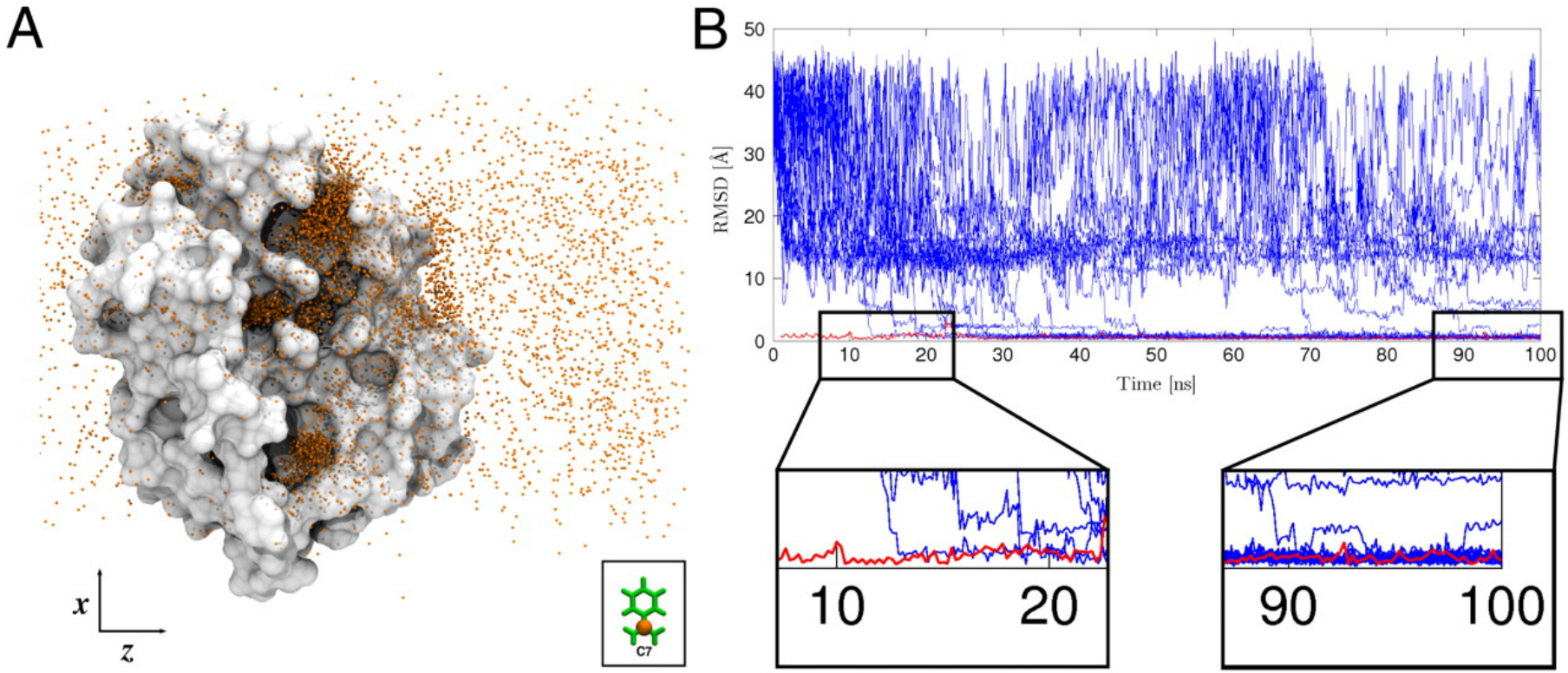
Free ligand binding has been used in the past to describe computational experiments in which, typically, a ligand is placed at a certain distance from the target protein and first by diffusion and

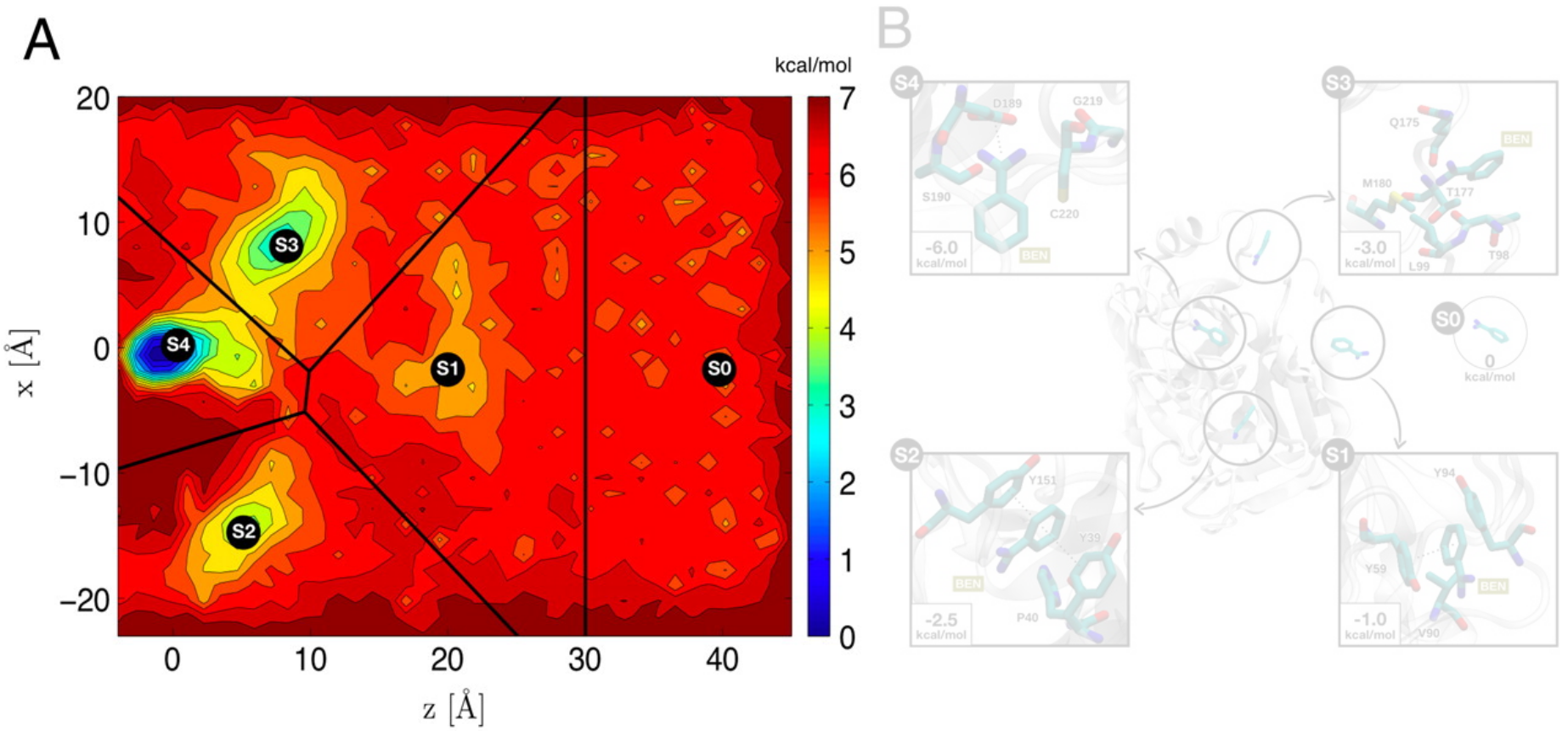
n the pro-





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.

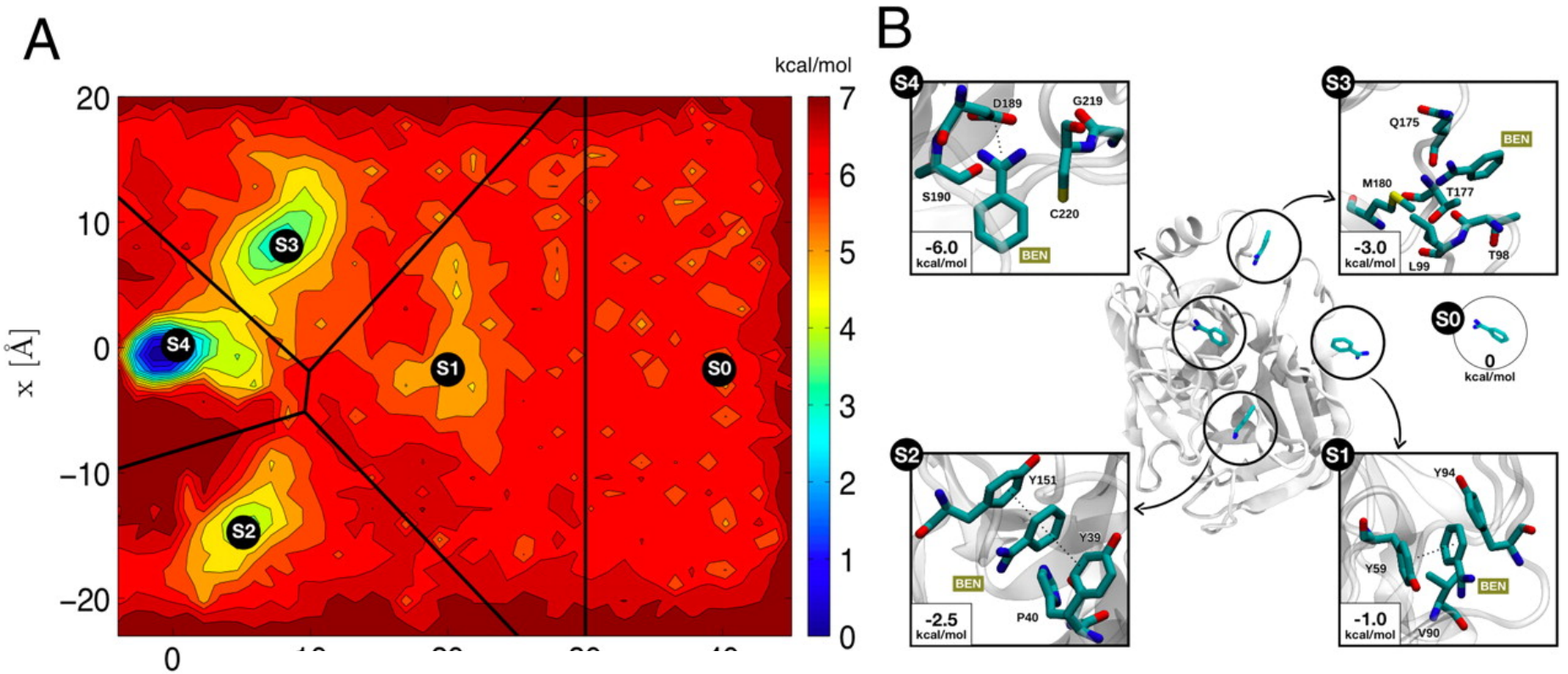


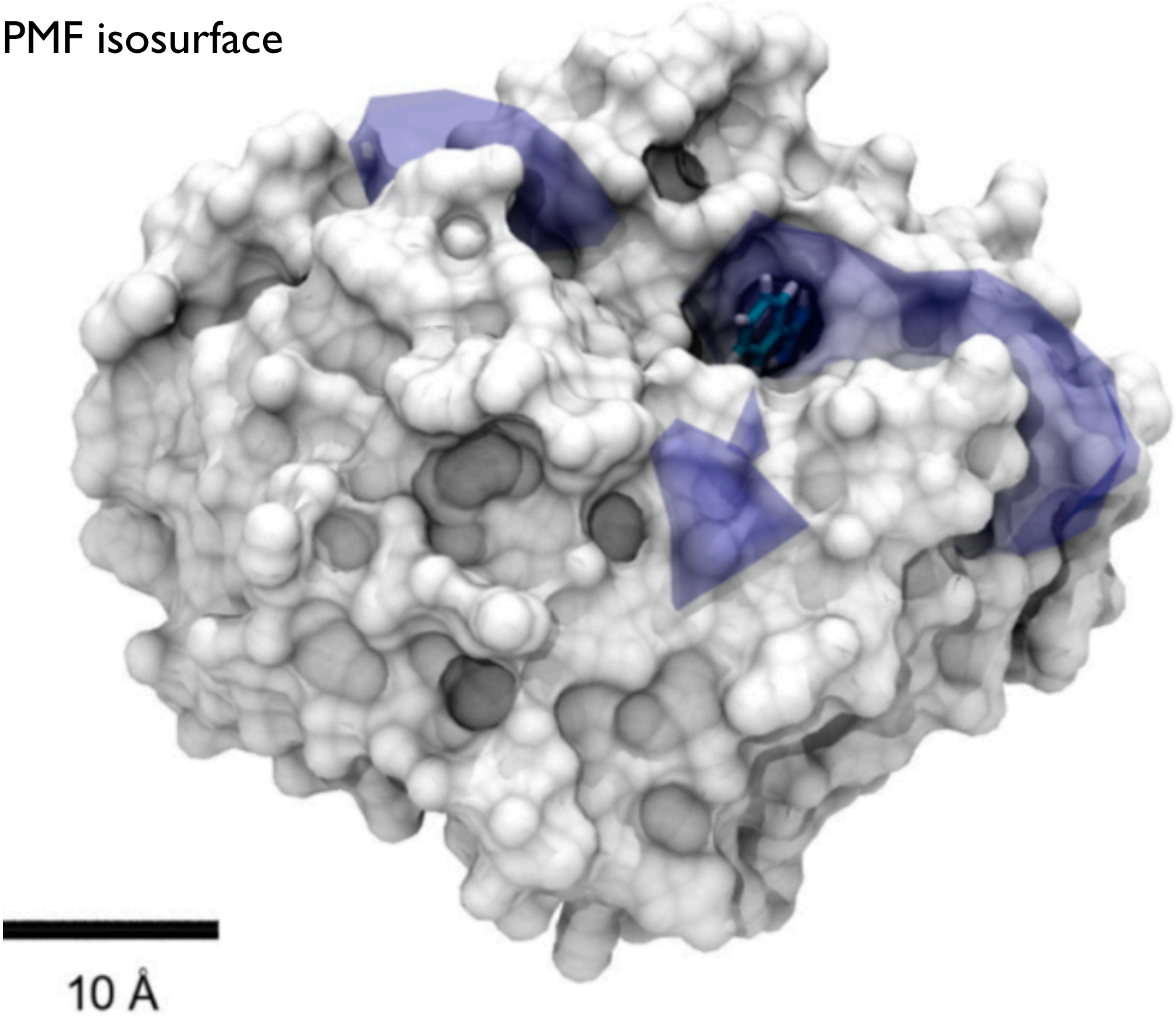
Table S1. Transition probabilities for the metastable states in the five-state coarse-grained model at a lag time of 50 ns

	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

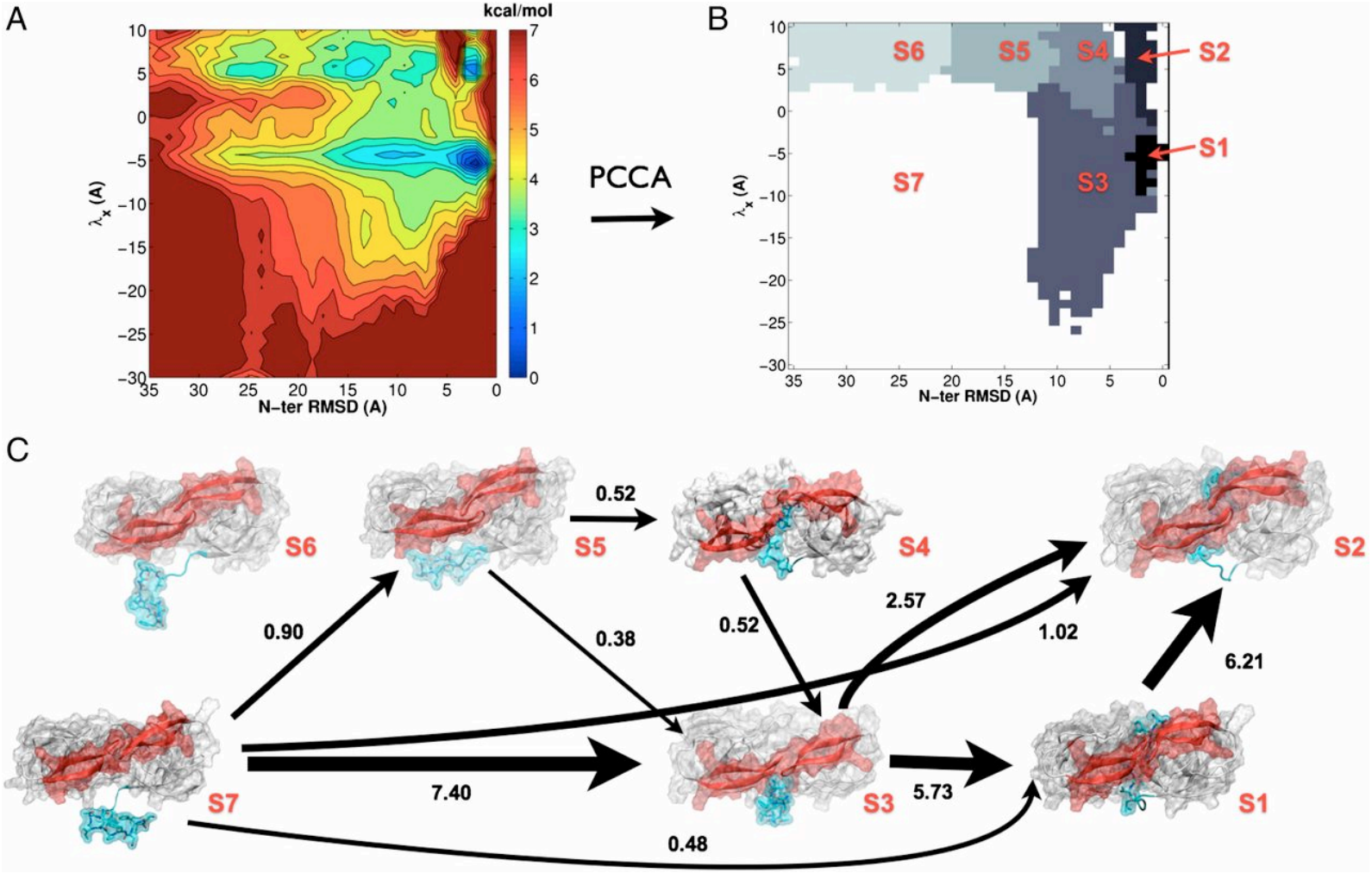
Identification of
identified from th
state S0 and the
from S3 from th
characterization
interactions with
of benzamidine (
between the Q17

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



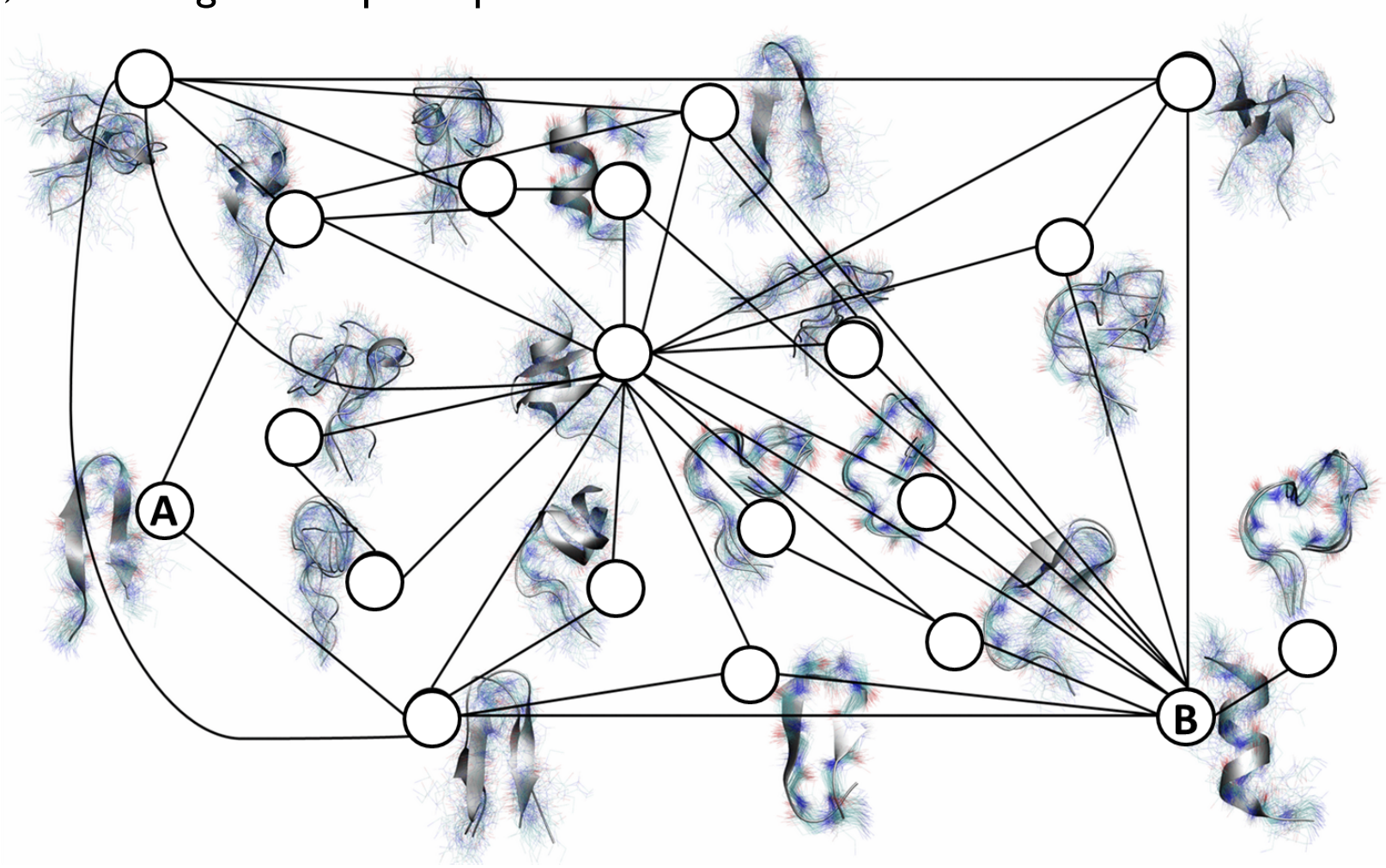
Kinetic characterization of the critical step in HIV-1 protease maturation



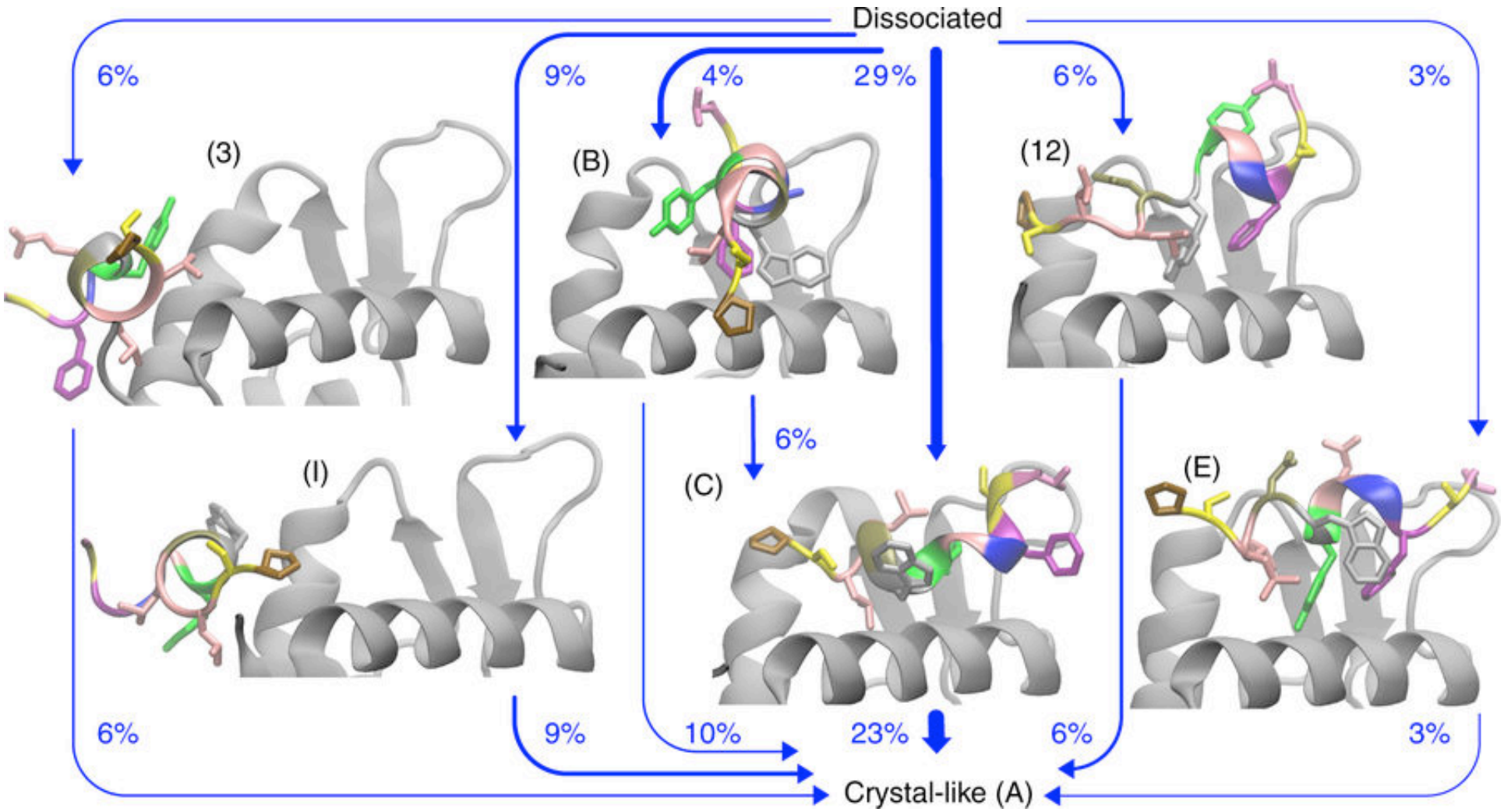
www.pnas.org/content/109/50/20449.full

Folding kinetics

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



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)