From Free-Energy Landscapes to Markov State Models: Towards a New Description of Allostery

Gregory Peters[1], Javier Estrada[2], John Biddle[2], and Jeremy Gunawardena[2]

[1]Department of Chemistry, Pacific Lutheran University, Tacoma, WA, 98447, U.S.A., [2]Department of Systems Biology, Harvard Medical School, Boston, MA, 02115, U.S.A.
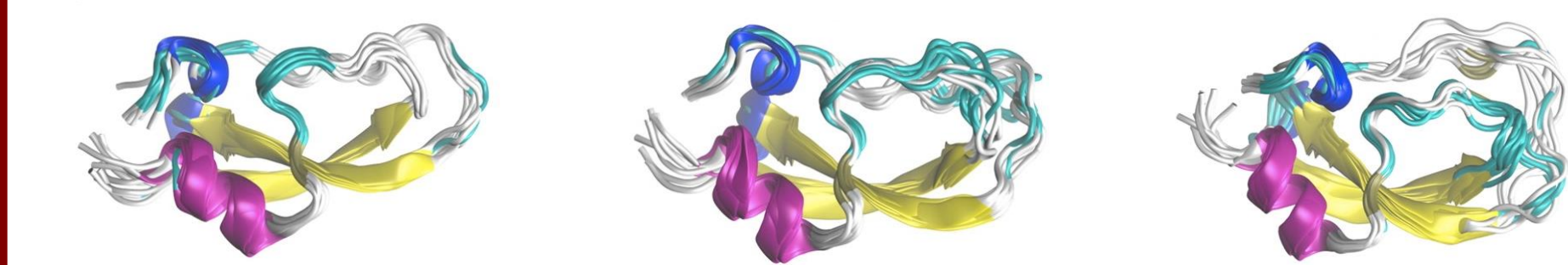
Introduction

Proteins are dynamic – they do not remain in a single, static conformation. Proteins exist in an ensemble of conformations, continually interconverting between conformations with varying energies. This ensemble of conformations has a certain distribution that can be described by the underlying free-energy landscape. This landscape exists in some extreme high-dimensional phase space of atomic motions. Often, the landscape is approximated in a highly-reduced view but in recent years it has been explored more directly through the use of molecular dynamics (MD) methods. A more discriminating approximation of the energy landscape is through the use of Markov state models (MSMs), which can approximate the free energy landscape by using the results of MD simulations. The states of a MSM are the free-energy minima, while the rates of transition between the states depend on the shape of the hills, and particularly the saddle points, surrounding the minima. From the resulting MSM, a protein conformation graph can be obtained.

To construct a MSM, the full continuous state space must be somehow discretized to obtain a tractable description of the dynamics. By performing a state space discretization, continuous states are grouped into discrete states, thus "erasing" information of the exact location within these states and "projecting" a continuous trajectory onto a discrete trajectory. The aim is to understand the process of transitioning from a free-energy landscape to a MSM and the resulting conformation graph. These conformation graphs provide the link required to begin formulating a description through the Gunawardena lab's linear framework.

The Linear Framework

The linear framework is a mathematical framework for doing time-scale separation in biochemical systems. It is based on graph theory and polynomial algebra. The framework allows for the elimination of the overwhelming molecular complexity found in cellular mechanisms such as allostery, post-translational modification, and gene regulation, and the construction of mathematical representations of how these mechanisms process information.
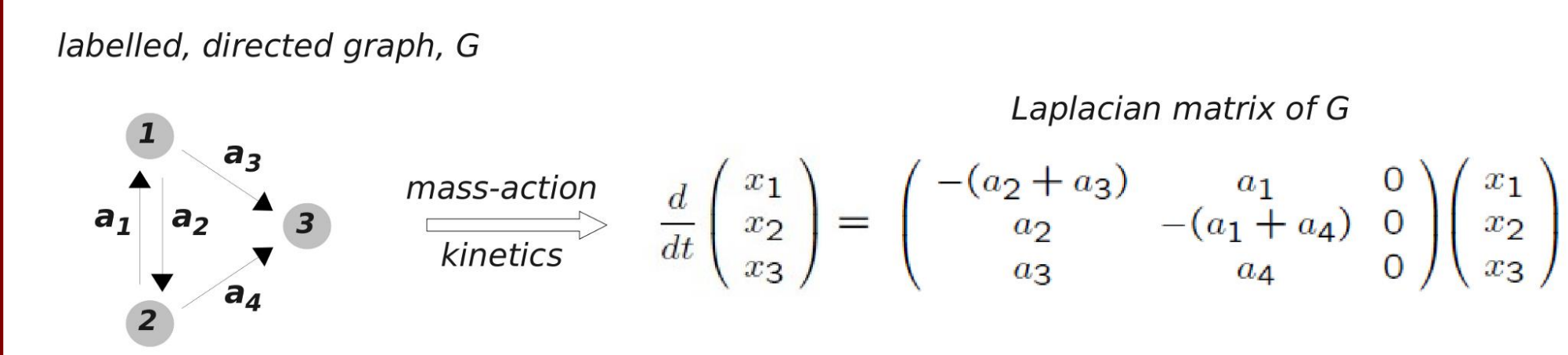
labelled, directed graph, G

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -(a_2+a_3) & a_1 & 0 \\ a_2 & -(a_1+a_4) & 0 \\ a_3 & a_4 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

Laplacian matrix of G

Figure 1: The Linear Framework
A labeled directed graph, *G*, gives rise to a system of linear differential equations by treating each edge as a first-order chemical reaction under mass-action kinetics, with the label as rate constants. The corresponding matrix is the Laplacian of *G*.[1]

Connecting the Linear Framework to Markov Models

Theorem[2] *Let X be any continuous time, finite-state space Markov process that is time homogenous, for which transition rates may be determined by:*

$$e_{ij} = \lim_{\Delta t \to 0} \frac{\Pr(X(t+\Delta t)=i \mid X(t)=j)}{\Delta t}.$$

The master equation of X is identical to Laplacian dynamics on the graph G_X:

$$\frac{dX(t)}{dt} = \mathcal{L}(G_X) \cdot X.$$
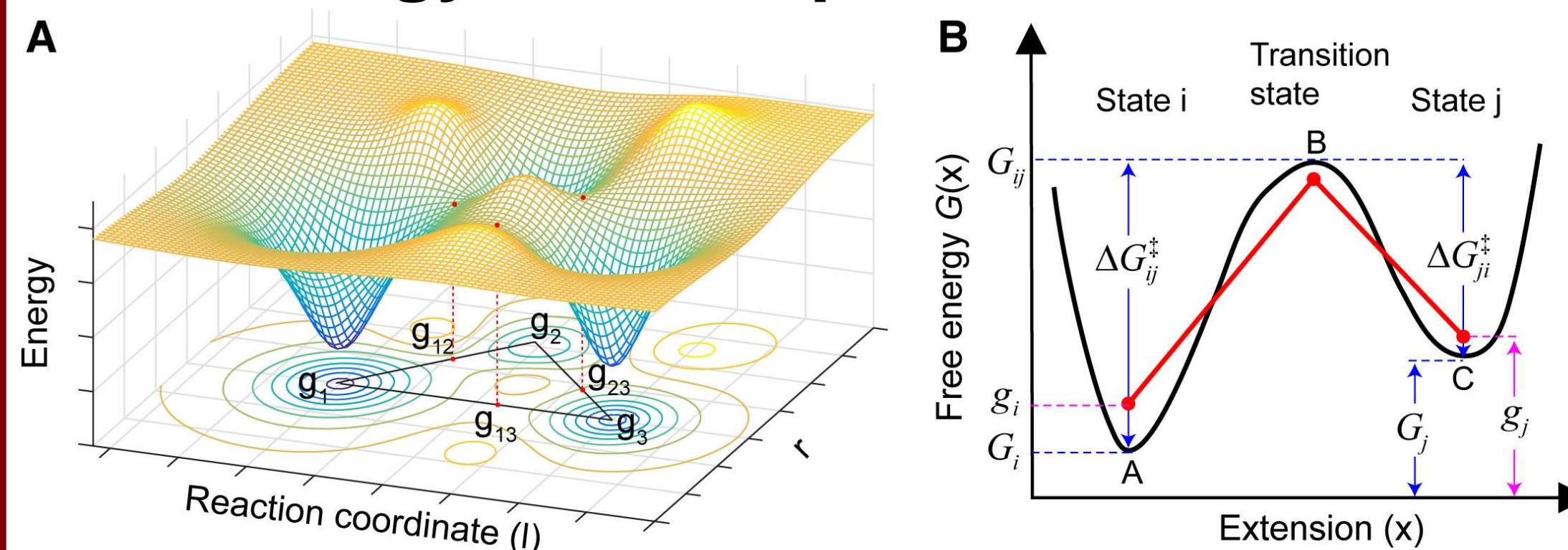
Free-Energy Landscape

Figure 2: Free-Energy Landscape of Protein Conformations
A. Schematic energy landscape for three-state protein conformations. B. Simplified energy landscape.[3]
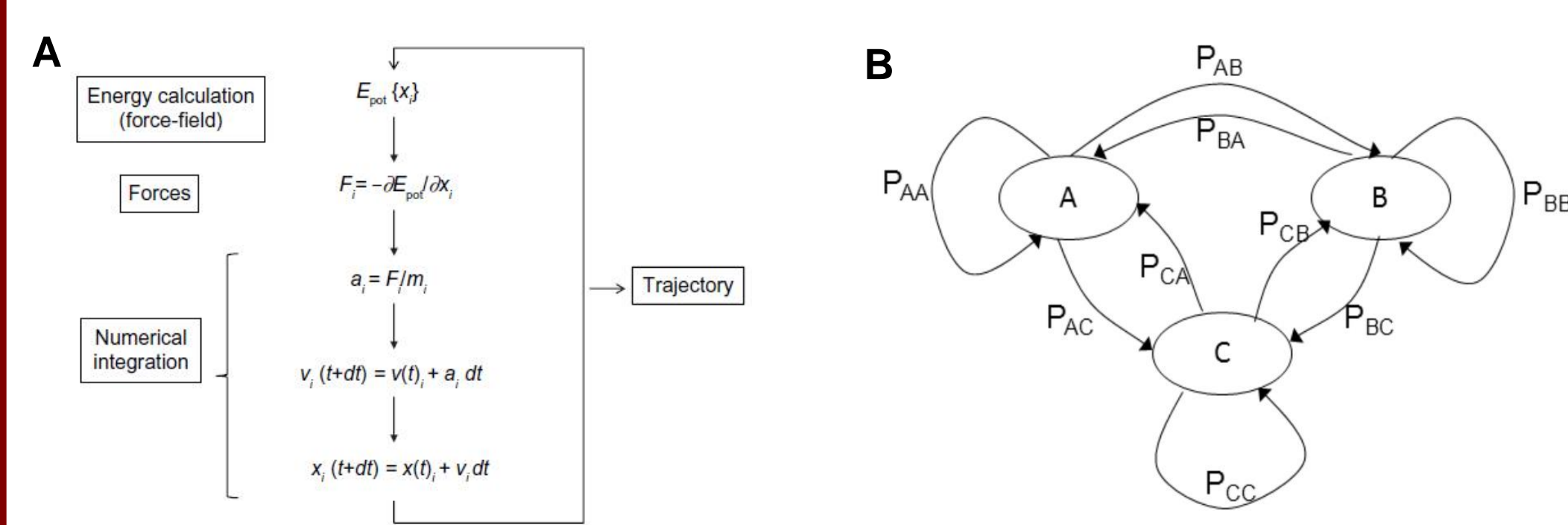
MD Simulation and Markov State Model

Figure 3: Example MD Algorithm and MSM
A. Basic MD algorithm.[4] B. Simple Markov state model.[5]
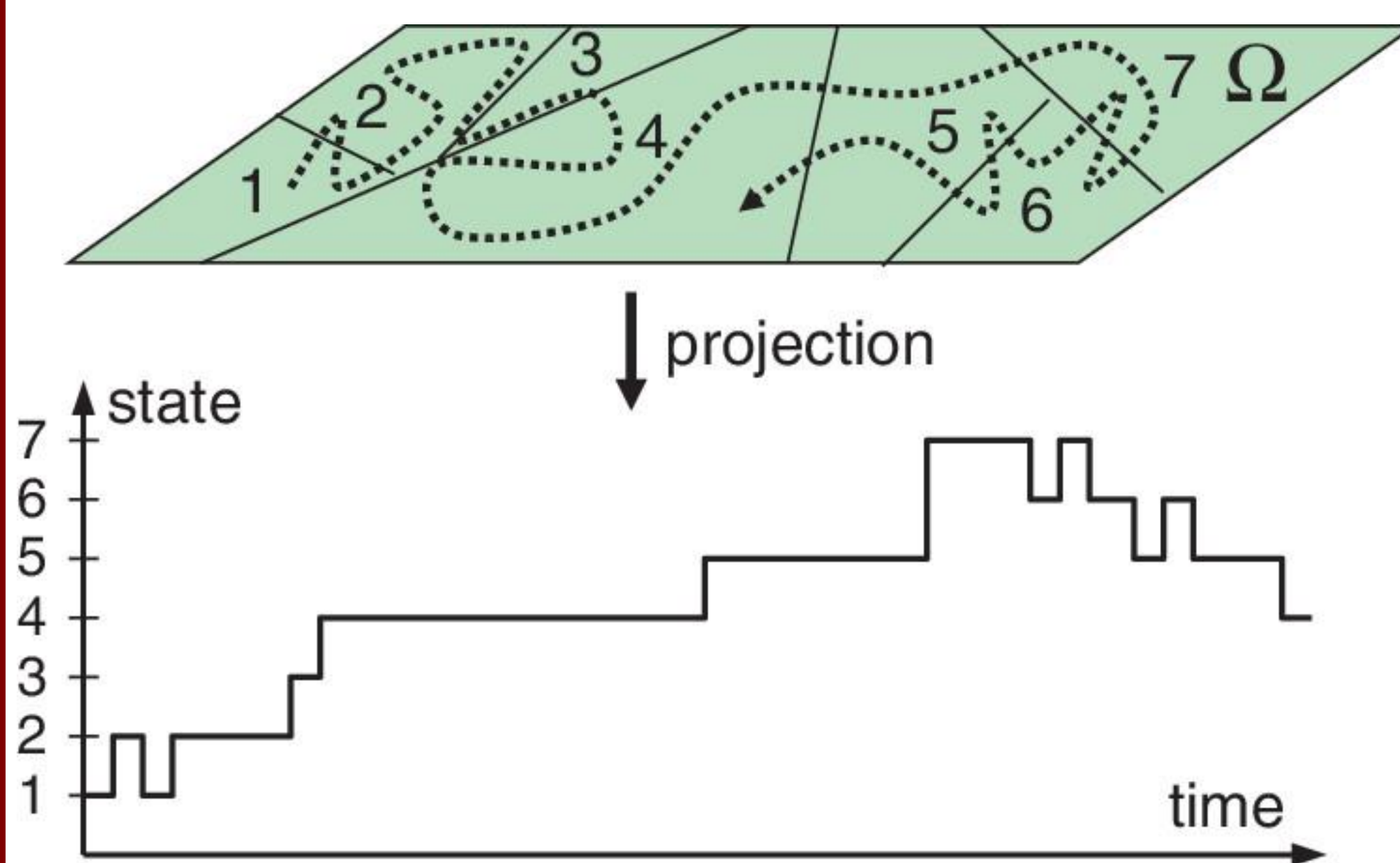
Dimension Reduction and Discretization

Figure 4: Discretization of State Space
The true continuous dynamics (dashed line) is projected onto the discrete state space.[6]
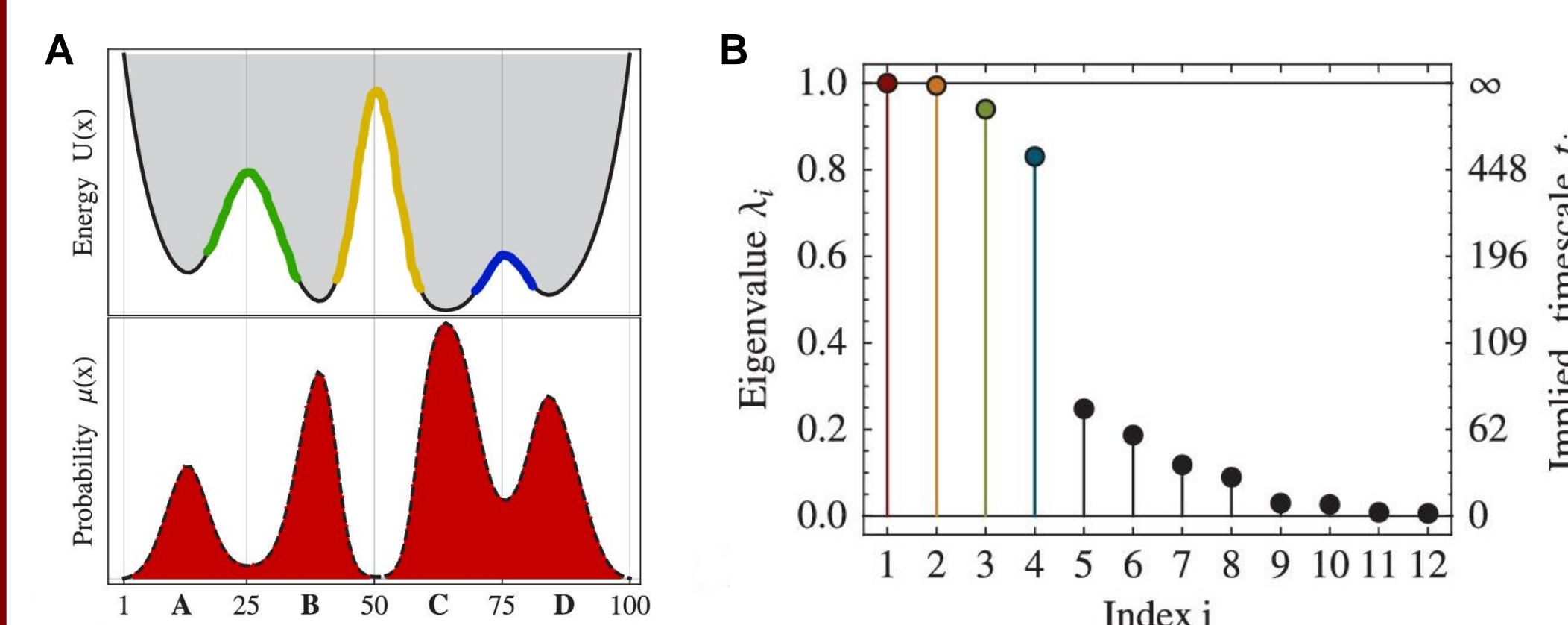
Figure 5: Identification of Slow Timescales
A. Potential energy function with four metastable states and corresponding stationary density. B. Eigenvalues of the transfer operator, the gap between the four metastable processes and the fast processes is clearly visible.[6]
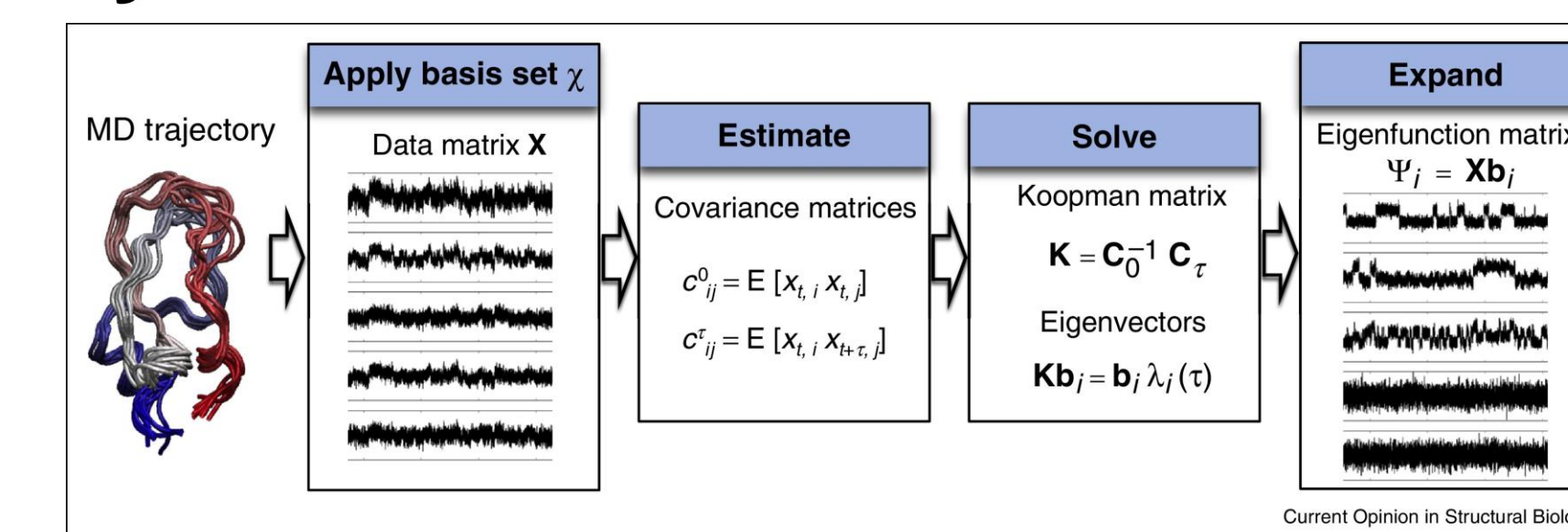
Time-Lagged Independent Component Analysis

Figure 6: tICA Algorithm
(1) Evaluate each basis function on all configurations sampled in a trajectory, (2) estimate covariance and time-lagged covariance matrix, (3) compute eigenvalues/eigenvectors, (4) project data matrix onto the eigenfunctions.[7]
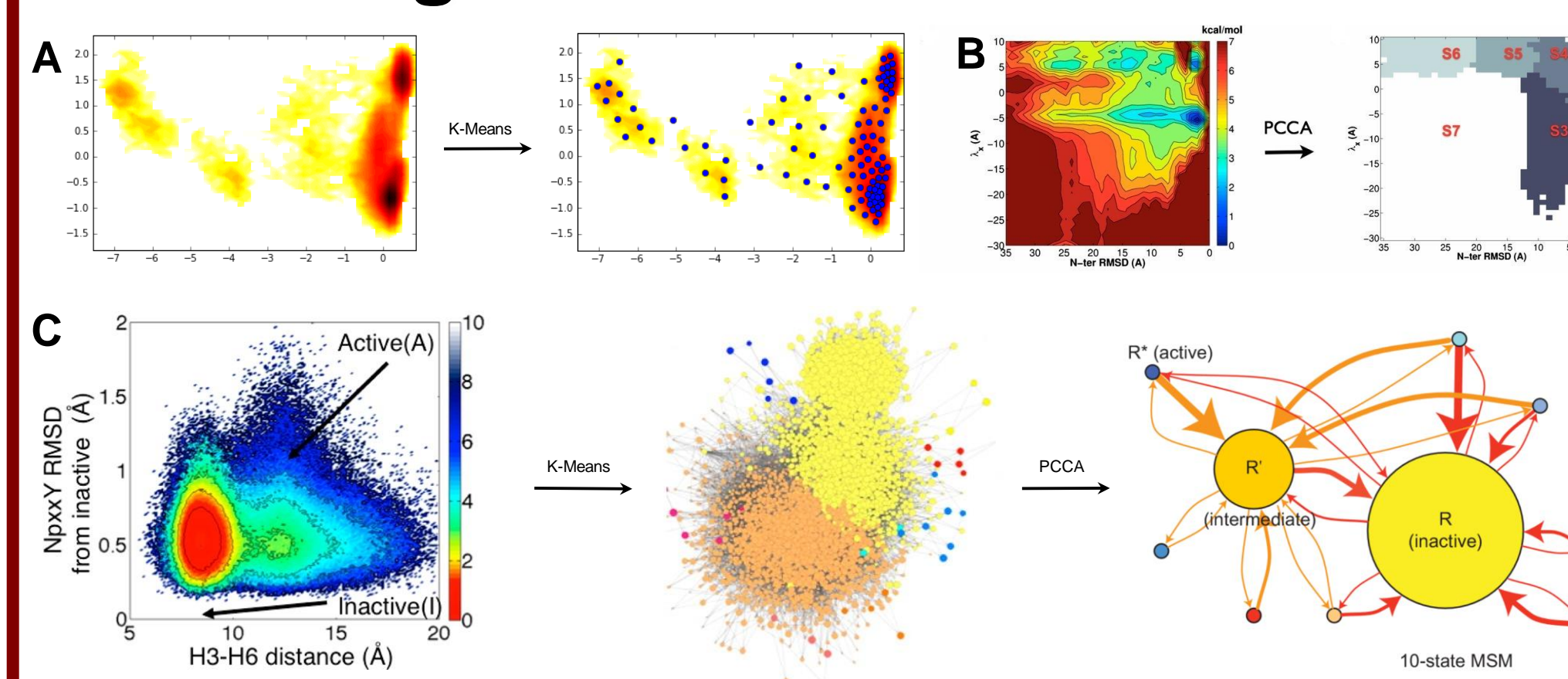
Clustering

Figure 7: Clustering Methods
A. Geometric clustering – k-means.[8] B. Kinetic clustering – Perron cluster analysis.[9] C. 3000-state model constructed using k-means and then coarse-grained to a 10-state model using PCCA.[10]
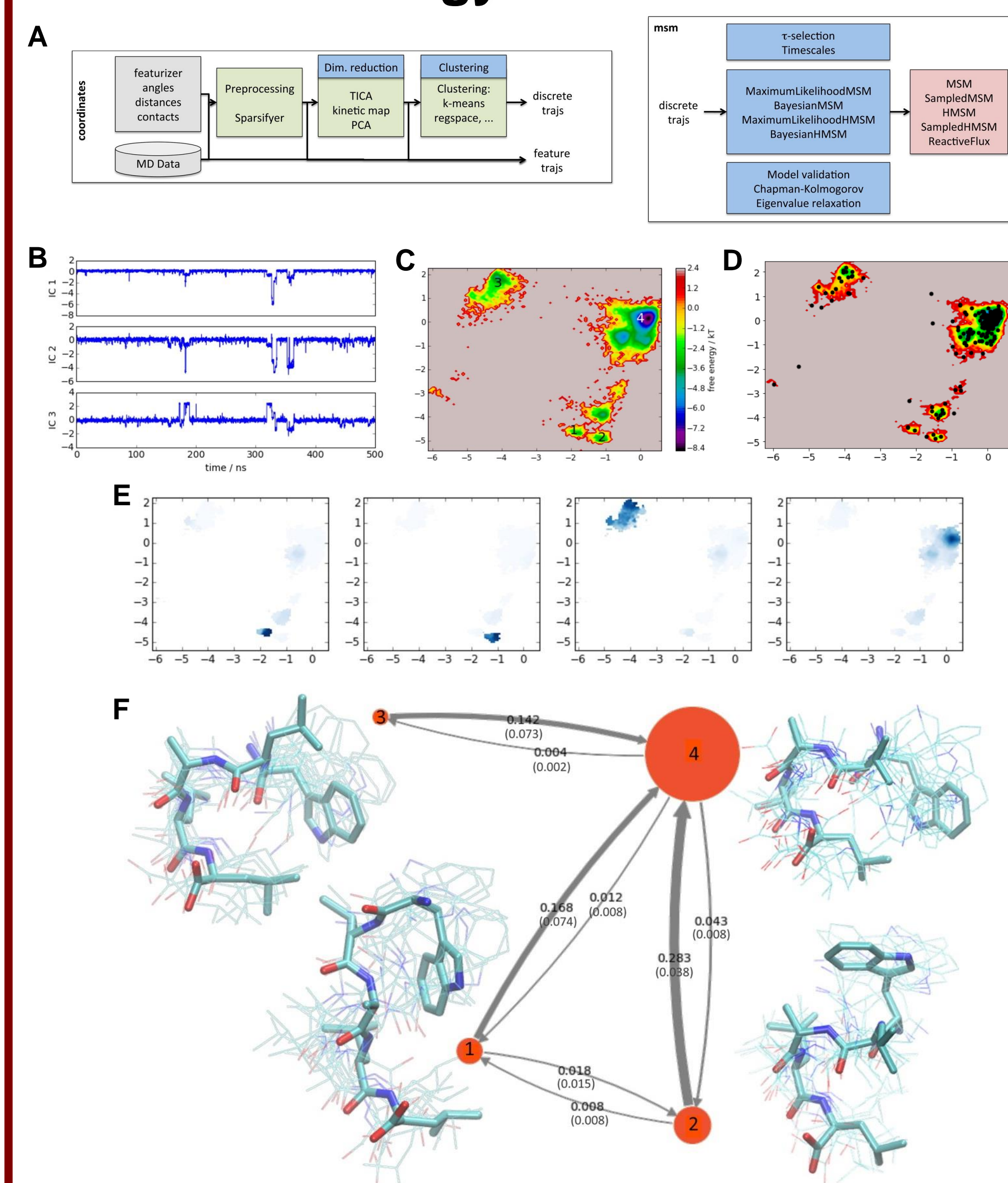
From Free Energy to Markov Model

Figure 8: Illustrative Markov State Model Analysis
A. PyEMMA software pipeline. B. Projection of a trajectory onto the slow collective coordinates (independent coordinates, ICs). C. Free energy landscape computed from a function of the two slowest ICs. D. Result of k-means clustering. E. Probability distributions for the four longest living metastable states determined through PCCA. F. Conformation graph obtained from a hidden Markov model based coarse-graining of the MSM. Rates are in ns⁻¹.[8]
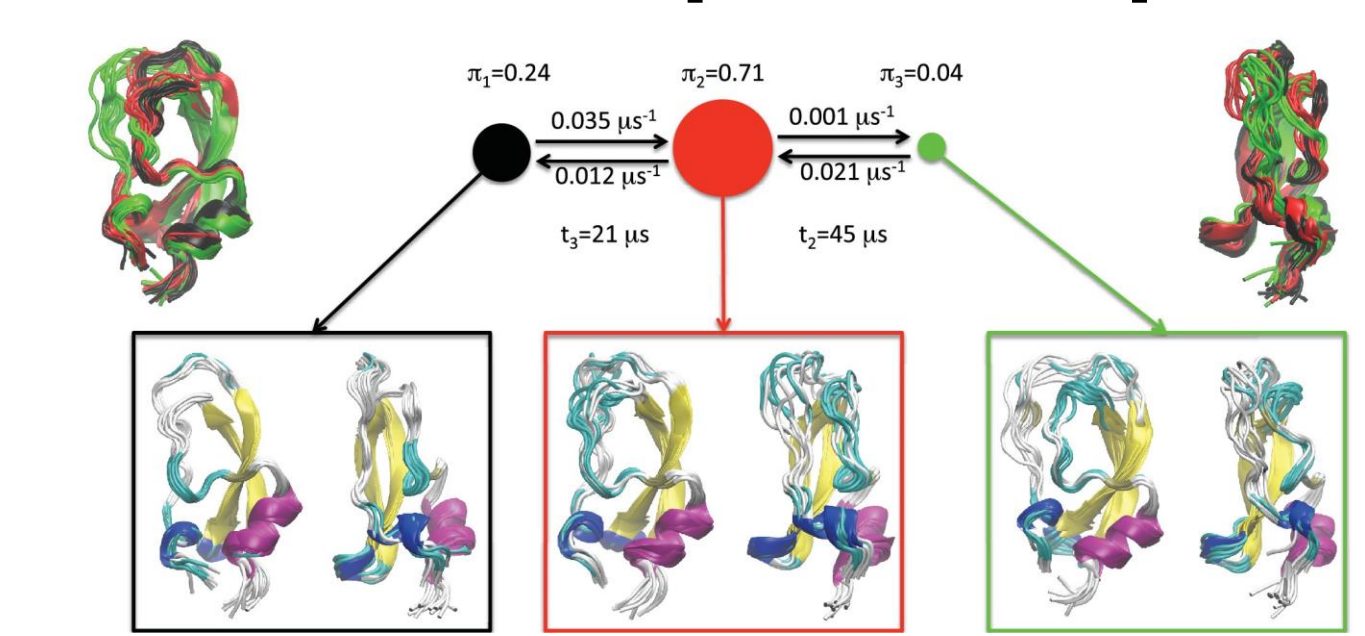
Conformation Graph Examples

Figure 9: Three-state conformation graph of bovine pancreatic trypsin inhibitor (BPTI) using a 1 ms simulation trajectory.[11]
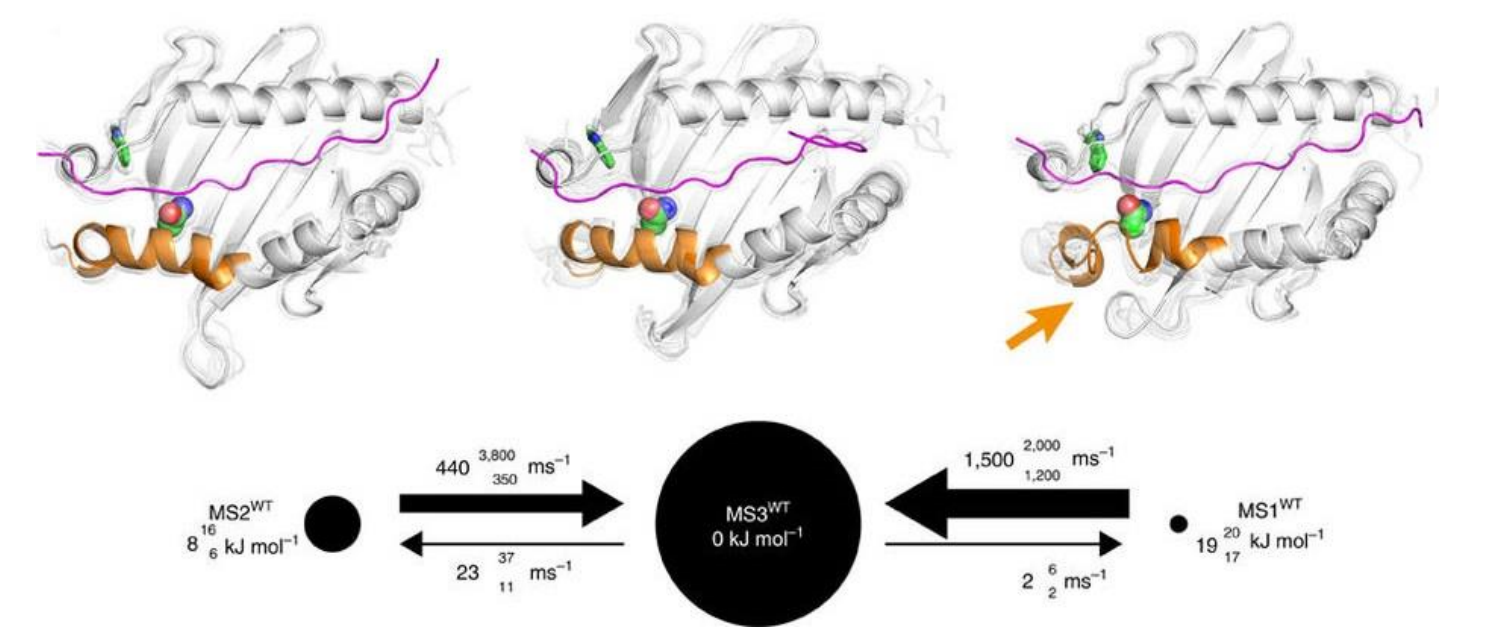
Figure 10: Three-state conformation graph of a peptide-major histocompatibility class II complex (pMHCII) using a 90 μs simulation trajectory.[12]
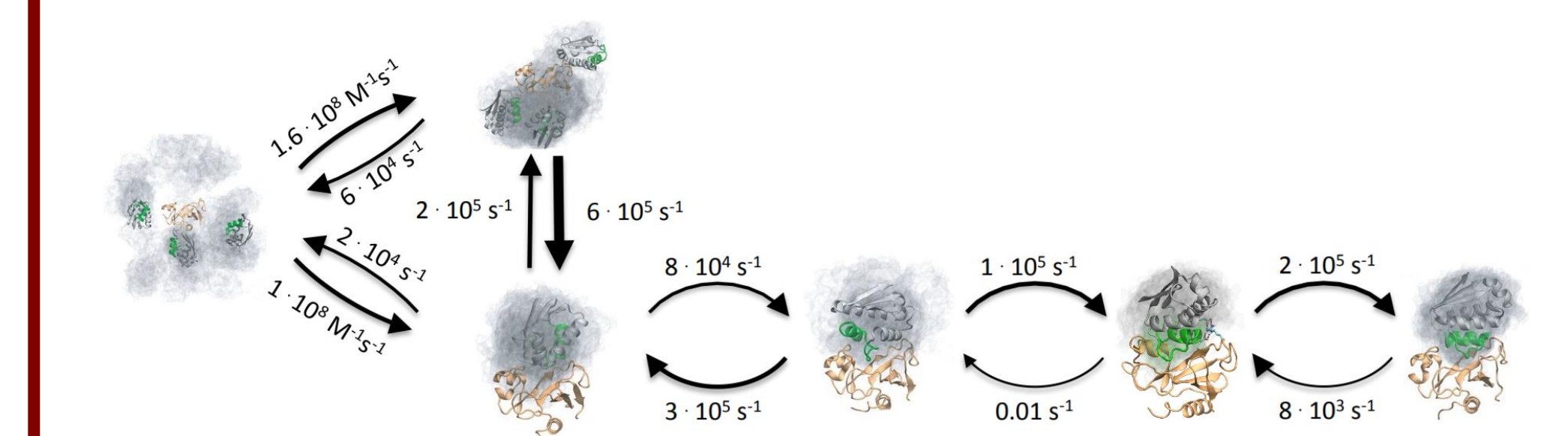
Figure 11: Coarse-grained conformation graph of the bacterial ribonuclease barnase and its inhibitor barstar using 2 ms of aggregate length simulation trajectories.[13]

Future Work

Just as protein conformational ensembles can be described by the free-energy landscape, so too can allostery. As its role in cellular information processing becomes clearer, the subject of allosteric regulation seems ripe for mathematical reconsideration. In trying to represent allostery in terms of a graph, we know the ratios of the rate constants but the individual values are harder to obtain. Because the ratios are the only parameters that matter at thermodynamic equilibrium, this suggests that a different kind of *equilibrium graph* will provide a more sparing representation. The main goal is to define an equilibrium *basic combined graph* in terms of a *conformation graph* and a *binding graph* and to derive a formula for allosteric regulation using coarse-graining.
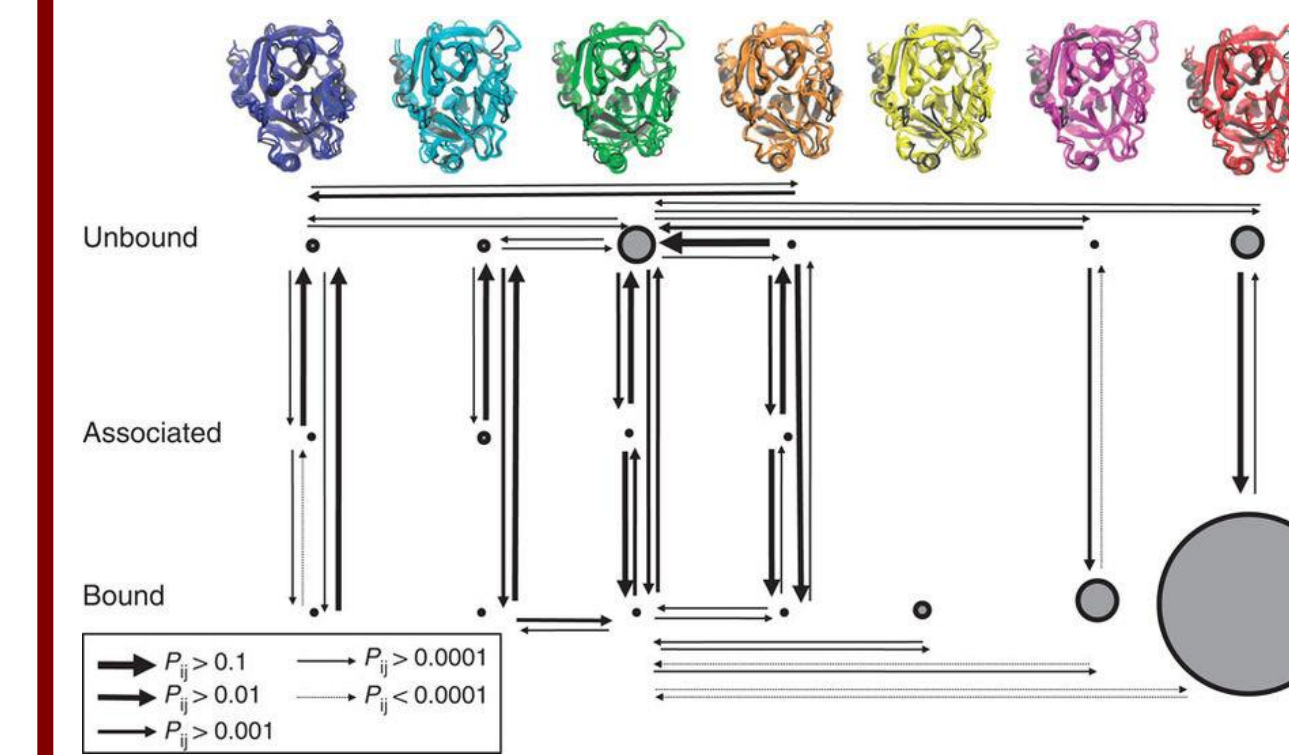
Figure 12: Basic Combined Graph Example:
Network of Trypsin-Benzamidine binding and Trypsin conformational dynamics.[14]

References

1. Gunawardena J. *PLoS ONE*, 7:e36321, 2012.
2. Mirazavi I and Gunawardena J. *Bull. Math. Biol.*, 75:2118, 49, 2013.
3. Zhang Y, Jiao J, Rebane, A. *Biophys. J.*, 111, 2110-2124, 2016.
4. Hospital A, Goñi JR, Orozco M, Gelpi JL. *Adv. Appl. Bioinform. Chem.*, 2015;8 37-47, 2015.
5. "MATLAB - A Fundamental Tool For Scientific Computing And Engineering Applications - Volume 2." Intechopen.com. N. p., 2017.
6. Prinz et al. *J. Chem. Phys.*, 134, 174105, 2011.
7. Noé F and Clementi C. *Curr. Opin. Struct. Biol.*, 43:141-147, 2017.
8. Scherer M, et al. *J. Chem. Theory Comput.*, 11, 5525-5542, 2015.
9. Sadiq S, Noé F, De Fabritiis G. *Proc. Natl. Acad. Sci. USA*, 109:50 20449-20454, 2012.
10. Kohlhoff K, et al. *Nat. Chem.*, 6, 15-21, 2013.
11. Noé F, Wu H, Prinz, J, Plattner N. *J. Chem. Phys.*, 139, 184114, 2013.
12. Wieczorek M, et al. *Nat. Commun.*, 7:13224, 2016.
13. Plattner N, Doerr S, De Fabritiis G, Noé F. *Nat. Chem.*, Advanced online publication, 2017.
14. Plattner N and Noé F. *Nat. Commun.*, 6:7653, 2015.