

Biomolecular simulations: from peptide dynamics to multi-protein assemblies

Gerhard Hummer

Laboratory of Chemical Physics

National Institute of Diabetes and Digestive and Kidney Diseases

National Institutes of Health

Bethesda, MD 20892-0520

U.S.A.

Department of Health and Human Services

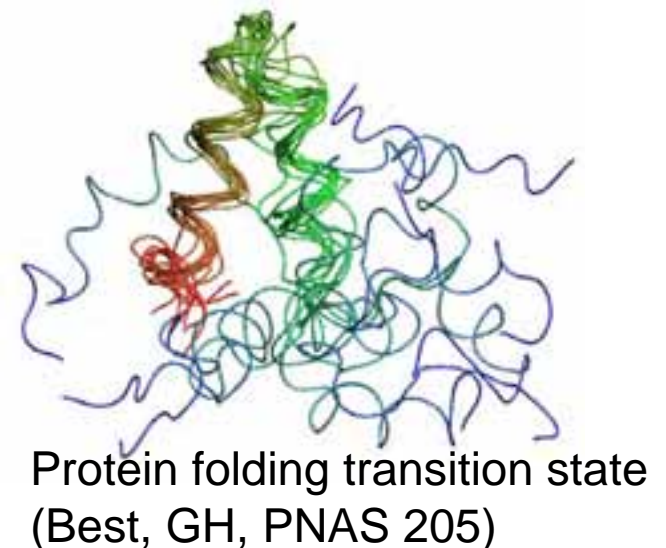
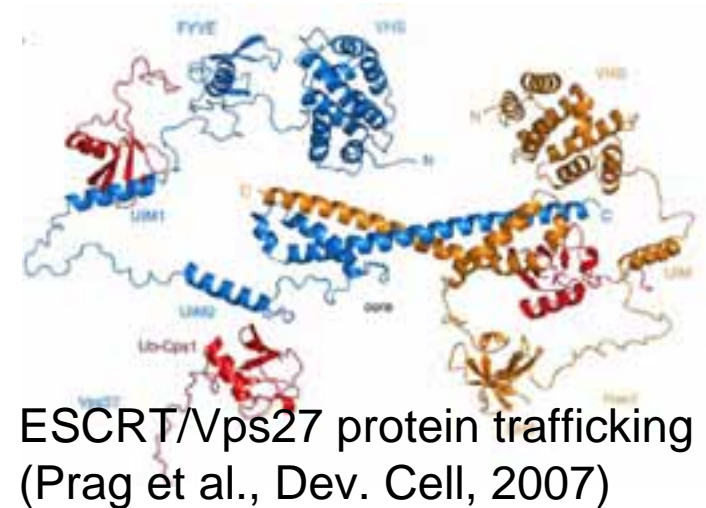


Lehigh University, 5 Oct 2009



Background: Molecular dynamics simulations in biology

- **Molecular dynamics (MD)**
 - Newtonian dynamics on approximate molecular energy surface
 - Structure, energetics, and dynamics of molecular systems
- **Resolution limits**
 - Space: ~ 1 Ångstrom (atom/bond)
 - Time: ~ 1 fs (collision/vibration)
- **Challenges**
 - Large length scales
 - from nm to $>\mu\text{m}$
 - Long time scales
 - from ps to $>\text{ms}$
 - Overwhelming detail



Current developments

- **Faster computers**
 - D.E.Shaw's "Anton" : millisecond simulations that would require ~ 100 years on regular hardware
- **Massive parallelization**
 - Folding@Home (V. Pande, Stanford): Molecular dynamics running on $>10^5$ screen savers around the world
- **Highly optimized code for supercomputers and clusters**
 - NAMD (Schulten, UIUC), Gromacs (Hess, Lindahl): atomistic simulation of large systems; $>10^6$ atoms (ribosome; Sanbonmatsu, Los Alamos)

Multiscaling through coarse graining in *space*

- **Reduced representation**
 - Integrating out small features of atomistic system (e.g., 1 amino-acid = 1 particle)
- **Spatial coarse-graining procedures**
 - Transferable energy functions (physics-based or data-base derived “statistical” interaction potentials)
 - Structure-based energy function (elastic network model, Gō model)
- **Challenge**
 - Small chemical features (e.g., hydrogen bonds) are closely connected to the overall structure and dynamics of biomolecular systems
- **Advantage**
 - Parameterization against relevant experiments helps avoid force-field issues of molecular dynamics simulations

Multiscaling through coarse graining in *time*

- Approximate representation of the *dynamics* of an atomically detailed model
 - Obtained by projecting full dynamics onto sub-space
 - Thermodynamics essentially exact (relative populations in configuration space)
 - Examples: master equation, Langevin dynamics, Brownian dynamics
- Challenges
 - Motions in proteins cover the spectrum from picoseconds (bond vibrations) to seconds (complete unfolding)
 - Motions are highly collective, making it difficult to identify relevant coordinates

Outline

1. Temporal coarse graining

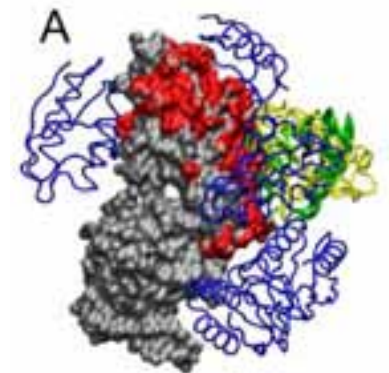
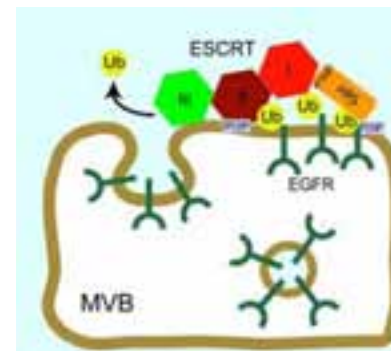
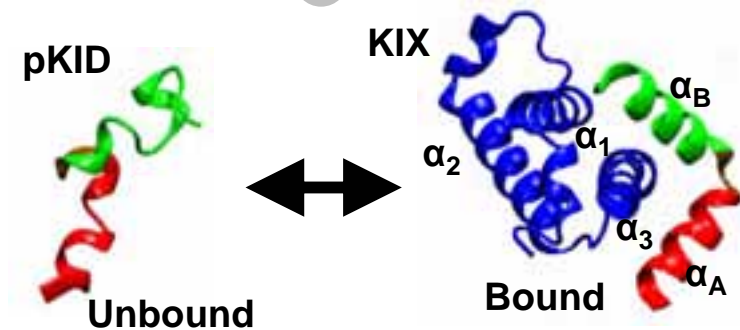
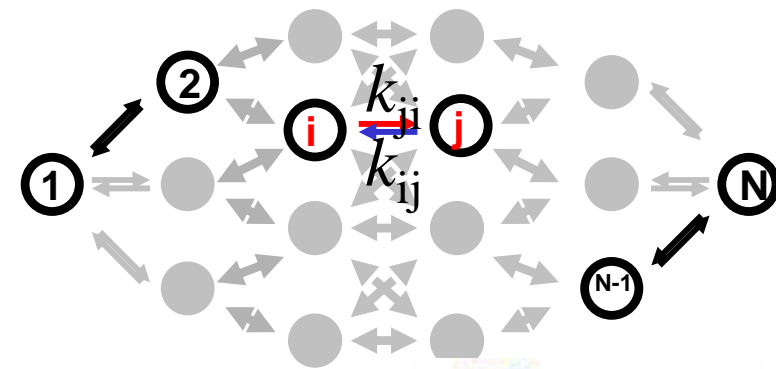
- Coarse master-equation for folding of a small helical peptide

2. Spatial coarse graining using structure-based potentials

- Binding of the natively unstructured transcription factor pKID/CREB to its co-activator KIX/CBP

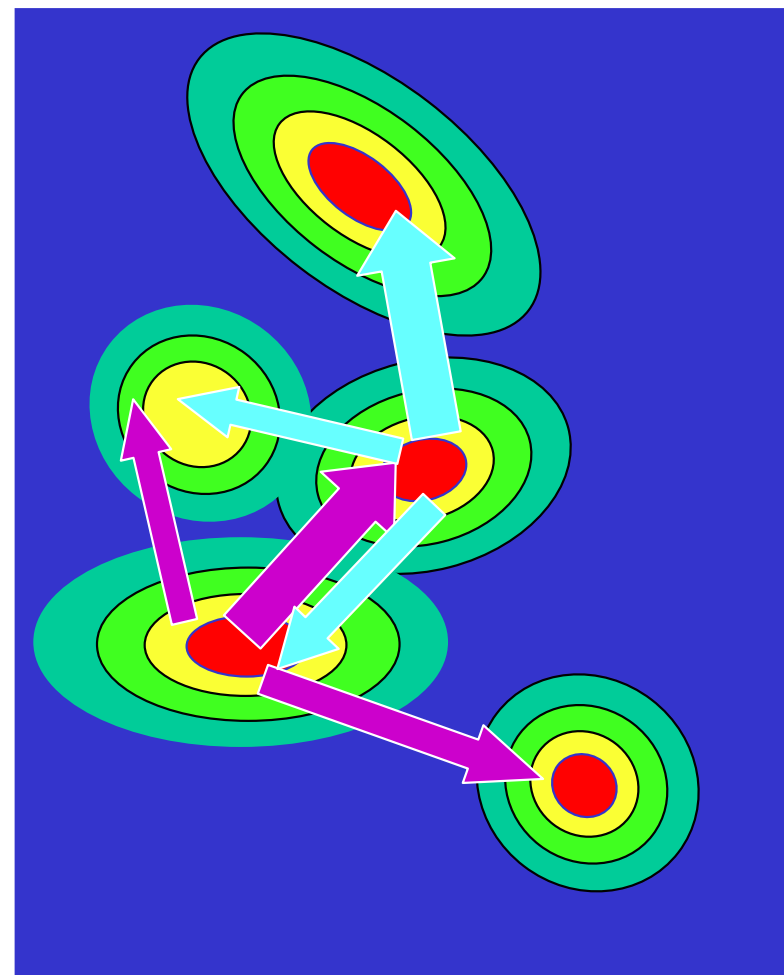
3. Transferable energy functions to study multi-protein assemblies

- Vps27/ESCRT protein sorting machinery
- Transient encounter complexes in protein-protein complex formation



1. Coarse graining in time

- *Circumvent* time-scale limitations of molecular dynamics simulations
 - Probe dynamics locally in configuration space
 - Combine local dynamics into global framework
 - Extrapolate to long-time dynamics, map free energy surfaces, calculate rate coefficients, identify transition states, ...



See also: Voter (parallel replica method)

1. Coarse master equation

(GH, Kevrekidis, *J. Chem. Phys.* **118**, 10762, 2003;

Sriraman, Kevrekidis, GH, *J. Phys. Chem. B* **109**, 6479, 2005;

Buchete, GH, *J. Phys. Chem. B* **112**, 6057, 2008 “Szabo Festschrift”)

- Configuration space divided into N cells

- Relative populations satisfy generalized master equation (Zwanzig, 1983)

$$\frac{dp_i(t)}{dt} = \int_0^t \sum_j R_{ij}(t-t') p_j(t') dt'$$

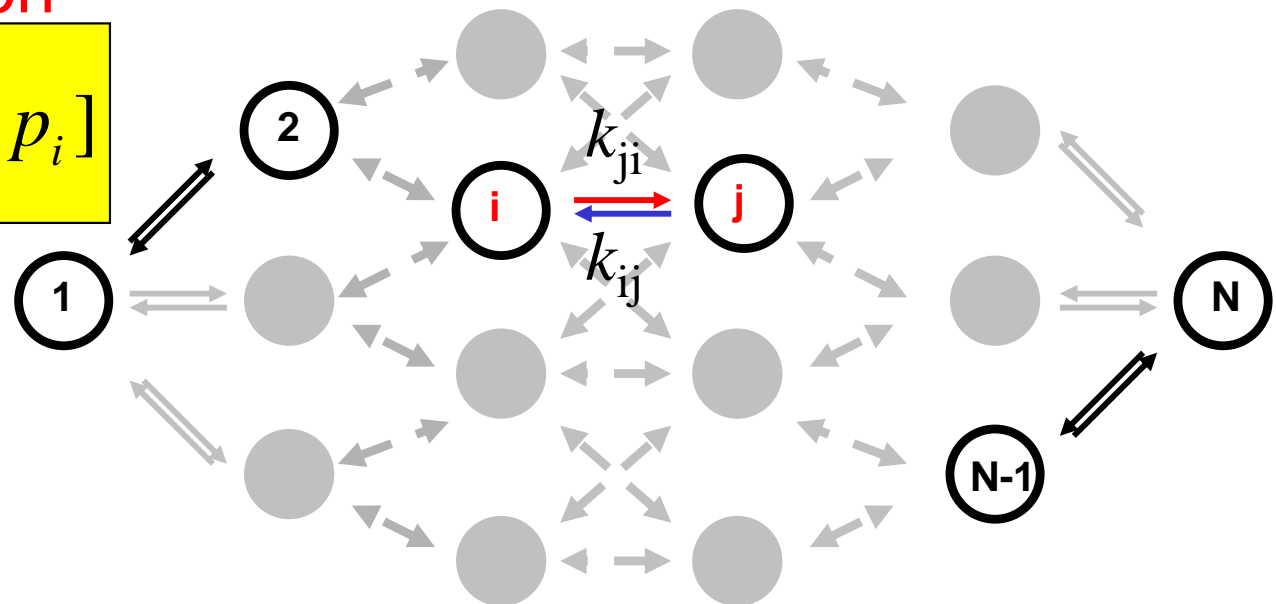
- If memory $R_{ij}(t)$ decays sufficiently fast, this reduces to the usual master equation

$$\frac{dp_i(t)}{dt} = \sum_j [k_{ij} p_j - k_{ji} p_i]$$

or in matrix form:

$$d\mathbf{p} / dt = \mathbf{K}\mathbf{p}$$

\mathbf{K} is rate matrix



See also: Schütte, Deuffhard, Swope, Noe, Pande, Levy, Karplus, Dinner, Elber, ...

Construction of master equation

1. Projection

- Assign instantaneous molecular structures along trajectory to states $s_\alpha = s(t_\alpha)$: (s_1, s_2, s_3, \dots) with $s_\alpha \in \{1, \dots, N\}$

2. Determine master equation that is “most consistent” with the observed projected trajectory

- Maximize likelihood (path action) with respect to rate coefficients of master equation

$$L = \prod_{\alpha} G(s_{\alpha+1}, \Delta t | s_{\alpha}, 0) \text{ where } G(i, \Delta t | j, 0) = [e^{\Delta t \mathbf{K}}]_{ij} \text{ is propagator (Green's function) in terms of matrix exponential}$$

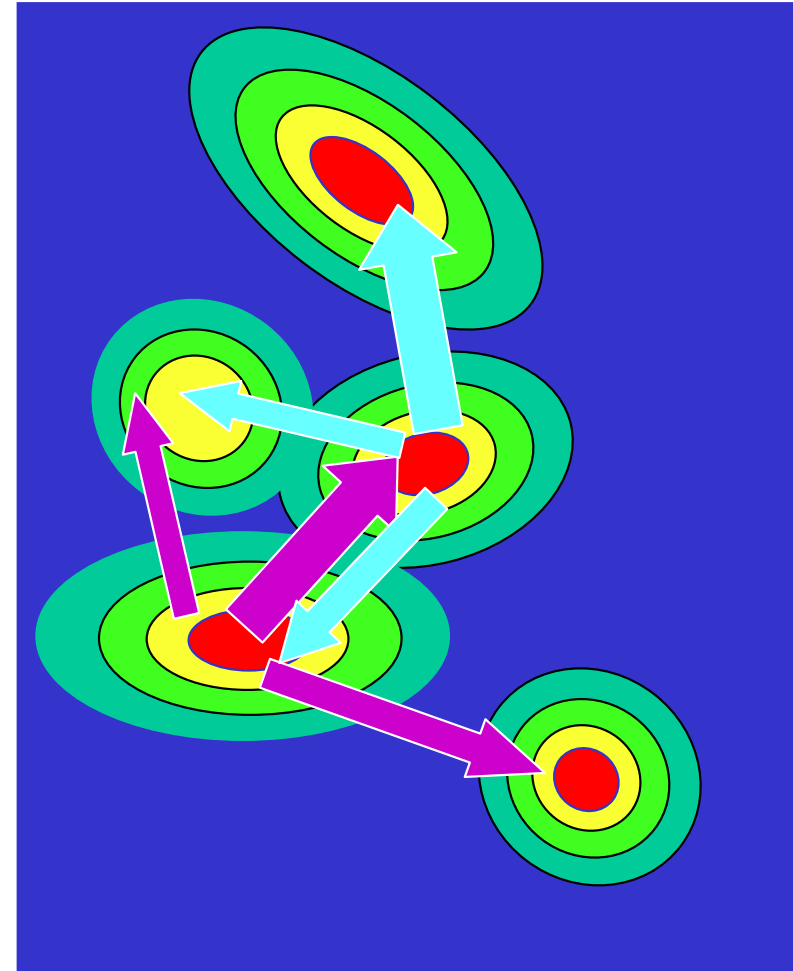
- $G(i, \Delta t | j, 0)$ is probability of being in cell i at time t , starting from cell j at time 0

3. Validate master equation

- Compare predicted and observed time correlation functions and populations

Local, “embarrassingly parallel” sampling

- *Dynamics can be sampled locally*
 - Run many “embarrassingly parallel” simulations to sample Green’s functions $G(i,t|j,0)$ from different starting points j
 - Combine local dynamics into global framework



Illustrative application: folding of a short helical peptide in water

(Buchete, GH, *J. Phys. Chem. B* **112**, 6057, 2008 "Szabo Festschrift")

- **Helix/coil transition of $\text{CH}_3\text{CO-Ala}_5\text{-NHCH}_3$**
 - AMBER-GSS force field (García & Sanbonmatsu, *PNAS*, 2002)
 - 1050 TIP3P water
 - 4x250 ns at 250, 300 and 350 K

- **Conformation space**

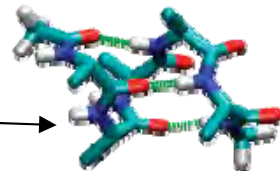
- 5 pairs of (Φ, Ψ) dihedral angles
- Each residue is either “helical” (1) or “non-helical” (0)
- $N=2^5=32$ possible conformations
- Relatively poor reaction coordinates (Bolhuis et al., *PNAS* 2000)

- **Binary notation**

- 00000: “all coil”

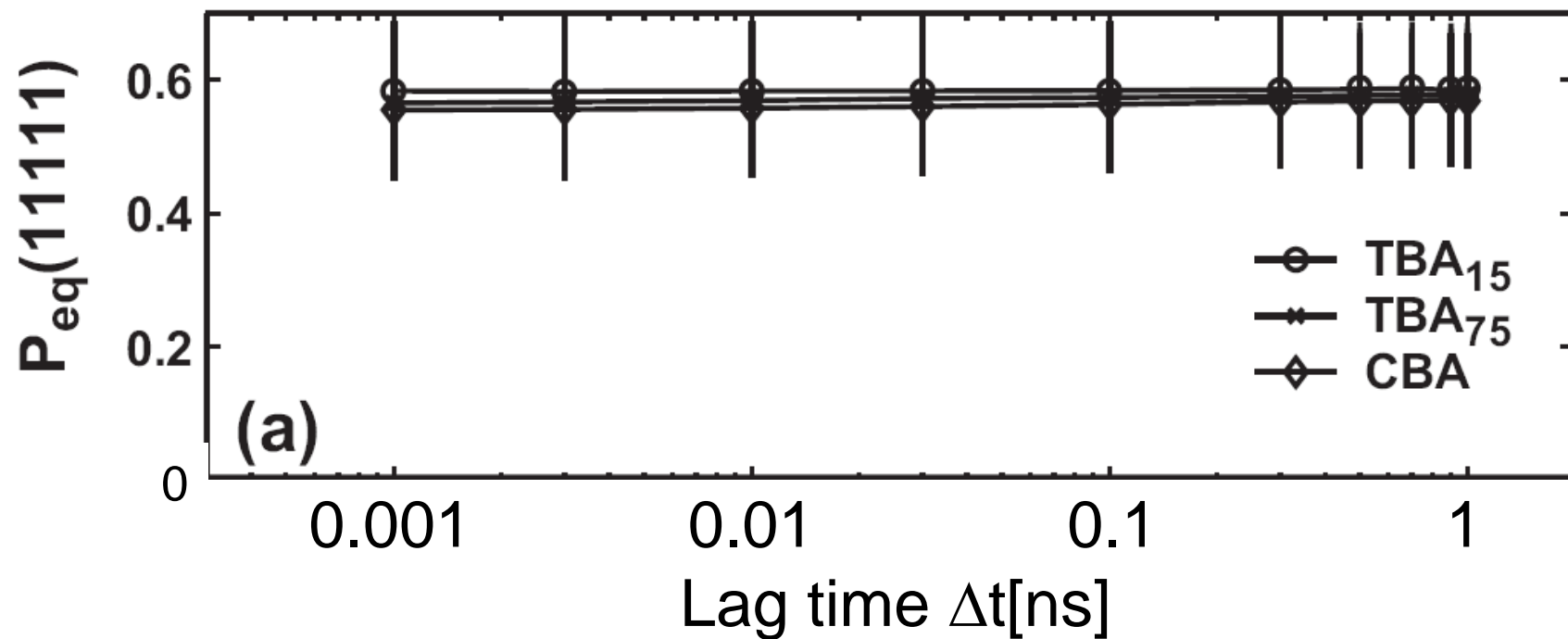


- 11111: “all helix”



Equilibrium populations independent of lag time and state assignment

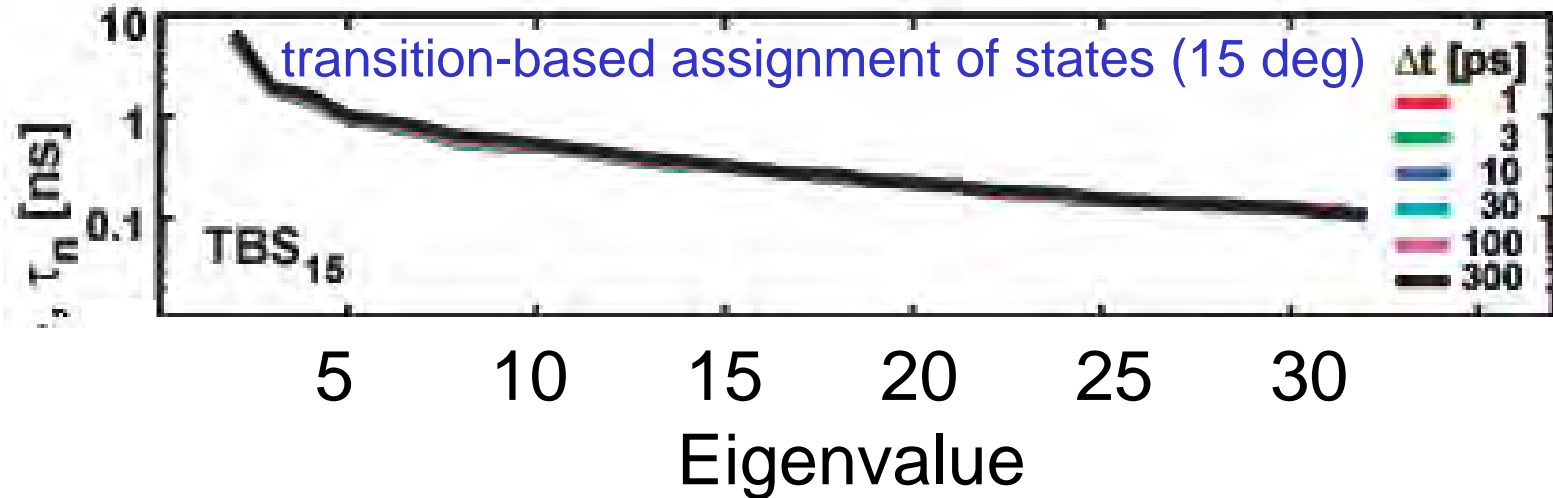
Population of all-helical state as a function of lag time Δt



$$L = \prod_{i,j=1}^N G(i, \Delta t | j, 0)^{N_{ij}}$$

Spectrum of rate matrix for different lag times

Relaxation times $\tau_n = -1/\lambda_n$ from the eigenvalues λ_n of \mathbf{K}



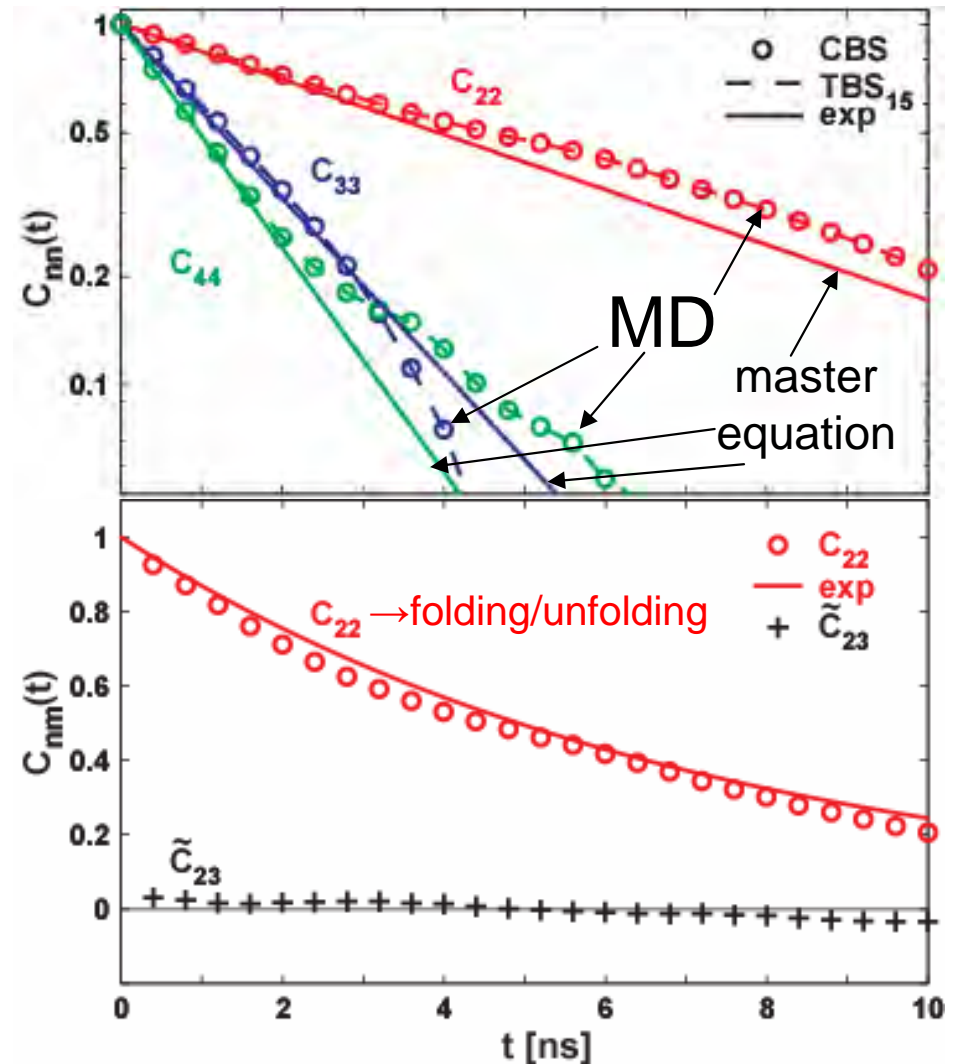
Validation: time-correlation functions of projected dynamics

- **Mode-correlation functions:**

- Trajectory projected onto the N states: $s(t) \in \{1, 2, \dots, N\}$
- Correlation function of left-hand eigenvectors should exhibit single-exponential decay:

$$\tilde{C}_{lm}(t) = \langle \psi_m^L[s(t)] \psi_l^L[s(0)] \rangle = \delta_{lm} e^{\lambda_m t}$$

- Excellent agreement of the entire spectrum, independent of state assignment
- Eigenmodes of master equation separate simulated dynamics into statistically independent Markov processes! \Rightarrow Identify macrostates: e.g., “folded”



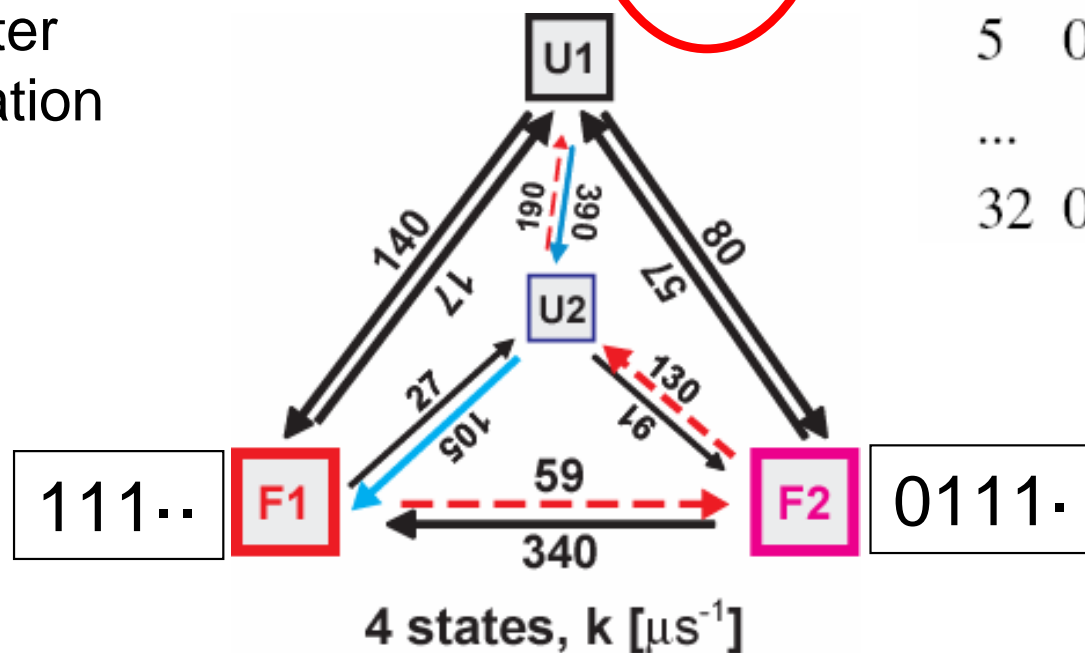
Validation of kinetic clustering: Amplitudes and relaxation times of 2 and 4-state representations

- Kinetic clustering**

- Microstates lumped based on eigenmodes of master equation

n	a_n	τ_n [ps]
Four-state projection		
2	0.929	7112.7
3	0.996	1774.5
4	0.978	1358.6

n	a_n	τ_n [ps]
32-state system		
2	0.964	7028.5
3	0.997	1985.7
4	0.982	1491.6
5	0.967	1174.1
...		
32	0.997	104.0



Conclusions (I): Coarse master equation

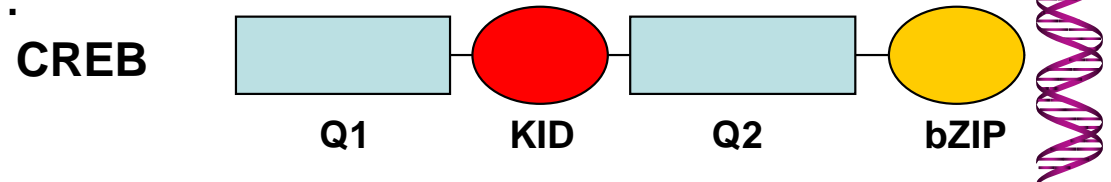
- Coarse master equation as a dynamics framework
 - long-time dynamics from short simulations
 - global free energy profiles from local kinetics
- Coarse master equation as an analysis tool
 - reduced representations of the microscopic system in terms of “macrostates” (such as “folded” and “unfolded”)
 - insights into molecular mechanisms through analytical transition states and reaction coordinates

2. Towards longer time scales: Binding of a natively unstructured transcription factor

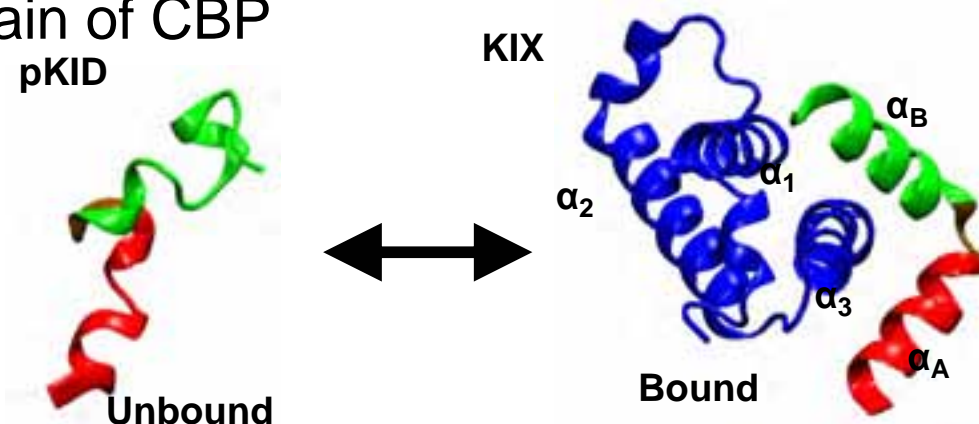
(Turjanski, Gutkind, Best, GH, *PLoS Comp. Biol.* **4**, e1000060, 2008)

- **Transcription factor cAMP response-element binding protein (CREB)**

- Involved in glucose homeostasis, cell survival and proliferation, learning and memory, ...



- Activated by phosphorylation of its kinase-inducible domain (KID) by protein kinase A (PKA).
 - Promotes recruitment of transcriptional co-activator CBP
- KID domain is natively unstructured, but becomes structured upon binding to KIX domain of CBP



Coupled folding and binding of pKID domain of CREB to KIX domain

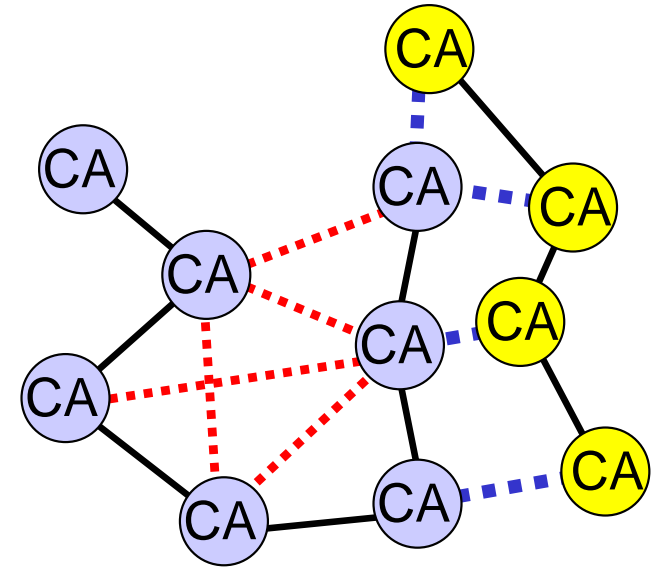
(Turjanski, Gutkind, Best, GH, *PLoS Comp. Biol.* 4, e1000060, 2008)

- Objective: identify mechanism of coupled folding and binding transition
 - Alternative mechanisms
 - “Pre-equilibrium/population shift:” peptide folds first and then binds
 - “Induced fit:” peptide binds unstructured and then folds
- Approach
 - Challenging system for atomistic molecular dynamics
 - Slow time scale: > 1 ms
 - Force-field issues: folding and binding of a peptide
 - Alternative: coarse graining using structure-based potentials

Spatial coarse graining: Native-topology (Gō) model of pKID/KIX binding

(Turjanski, Gutkind, Best, Hummer, *PLoS Comp. Biol.* **4**, e1000060, 2008)

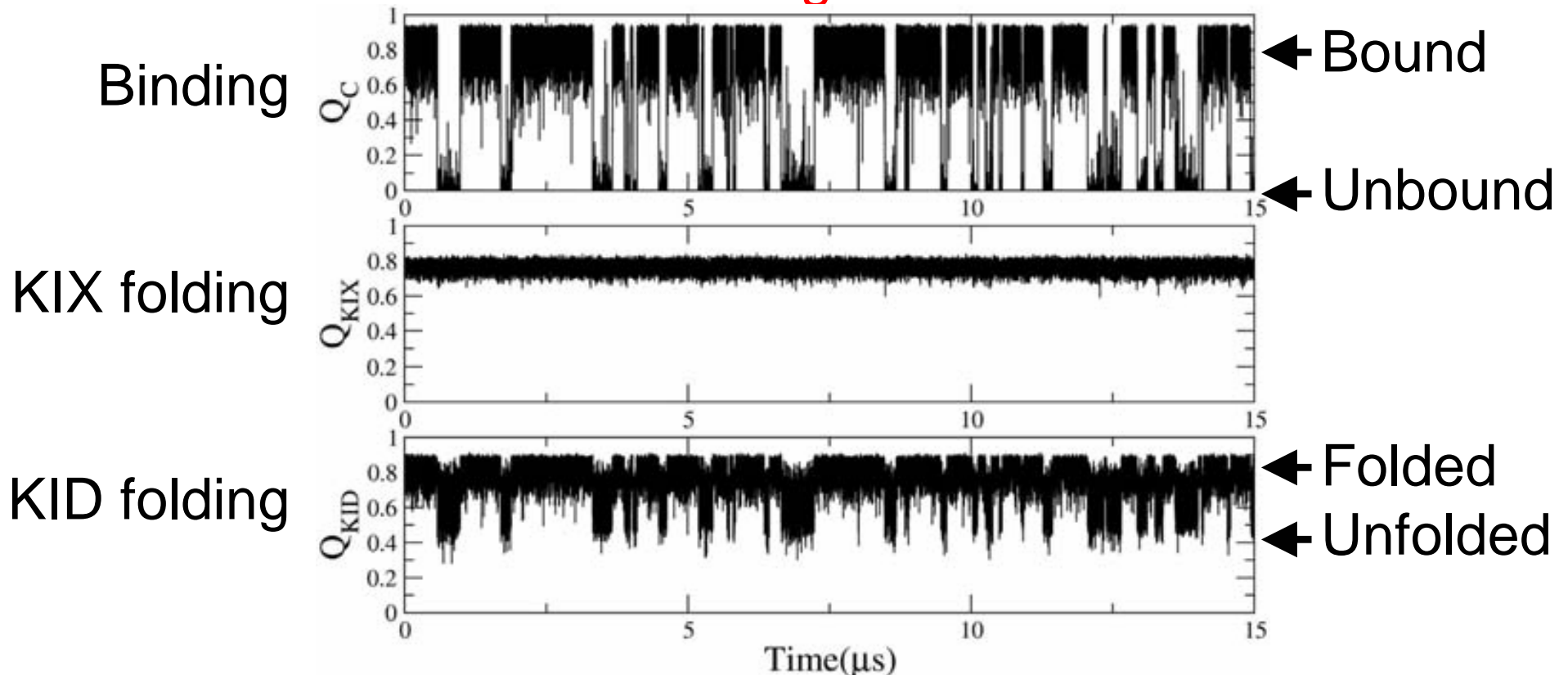
- pKID and KIX peptide chains represented as polymers with one bead per amino acid
 - Attractive residue interactions for native intra-peptide contacts
 - Attractive interactions for native inter-peptide contacts
- Effects of non-native attractive interactions probed by adding statistical contact potentials (Miyazawa, Jernigan, *J. Mol. Biol.* **256**, 623, 1996)
- Langevin dynamics at low solvent friction
 - Time scaled by factor ~ 1000 : 0.1 ps^{-1} versus $\sim 100 \text{ ps}^{-1}$ for water
 - Effective simulation time: $\sim 200 \text{ ms}$



Folding and binding

(Turjanski, Gutkind, Best, GH, *PLoS Comp. Biol.* 4, e1000060, 2008)

- Fractions Q_{KIX} and Q_{KID} of inter-peptide amino-acid contacts indicate folding
- Fraction Q_C of native inter-peptide amino-acid contacts indicates binding

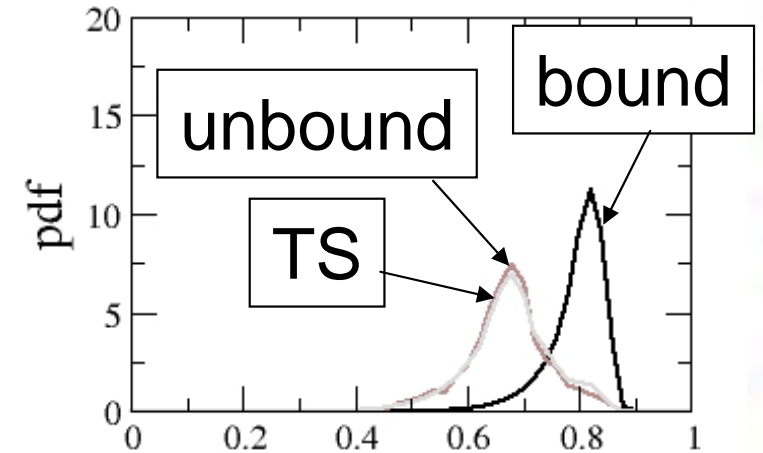
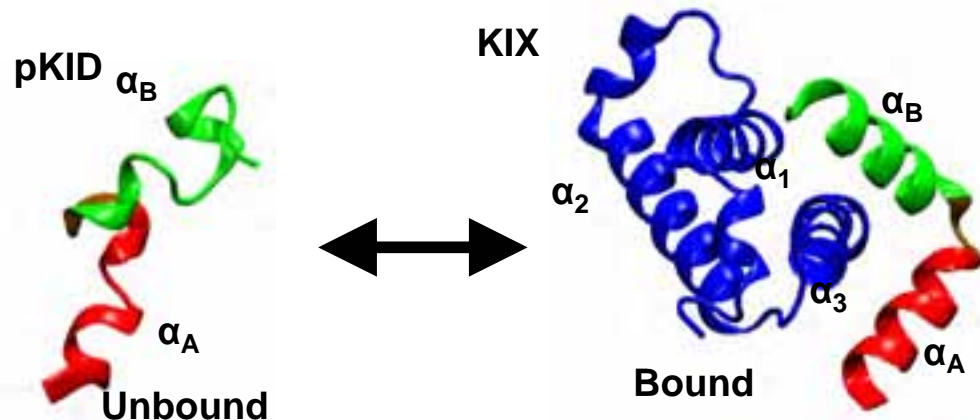


⇒ Folding of pKID and binding to KIX are strongly coupled

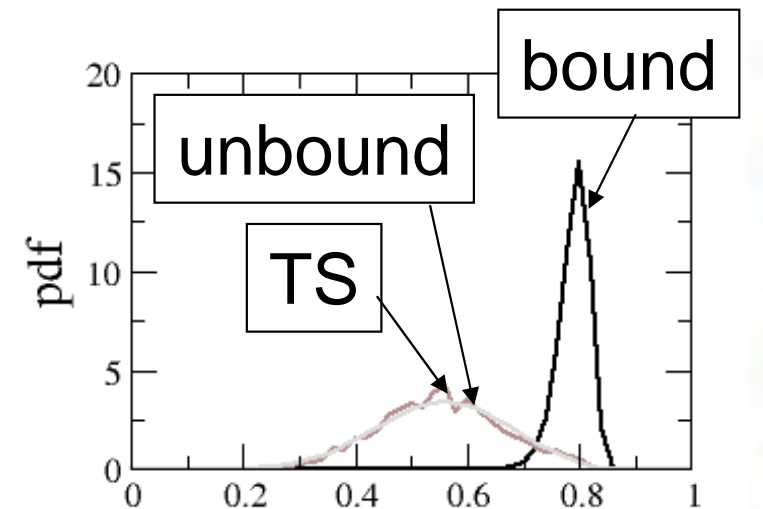
Transition state ensemble

(Turjanski, Gutkind, Best, GH, *PLoS Comp. Biol.* 4, e1000060, 2008)

- **Binding precedes folding**
 - At the transition state of the coupled folding/binding reaction, both α_A and α_B are unstructured
 - Intra-peptide contacts Q_A and Q_B are distributed as in the unbound state



Fraction Q_A of α_A contacts



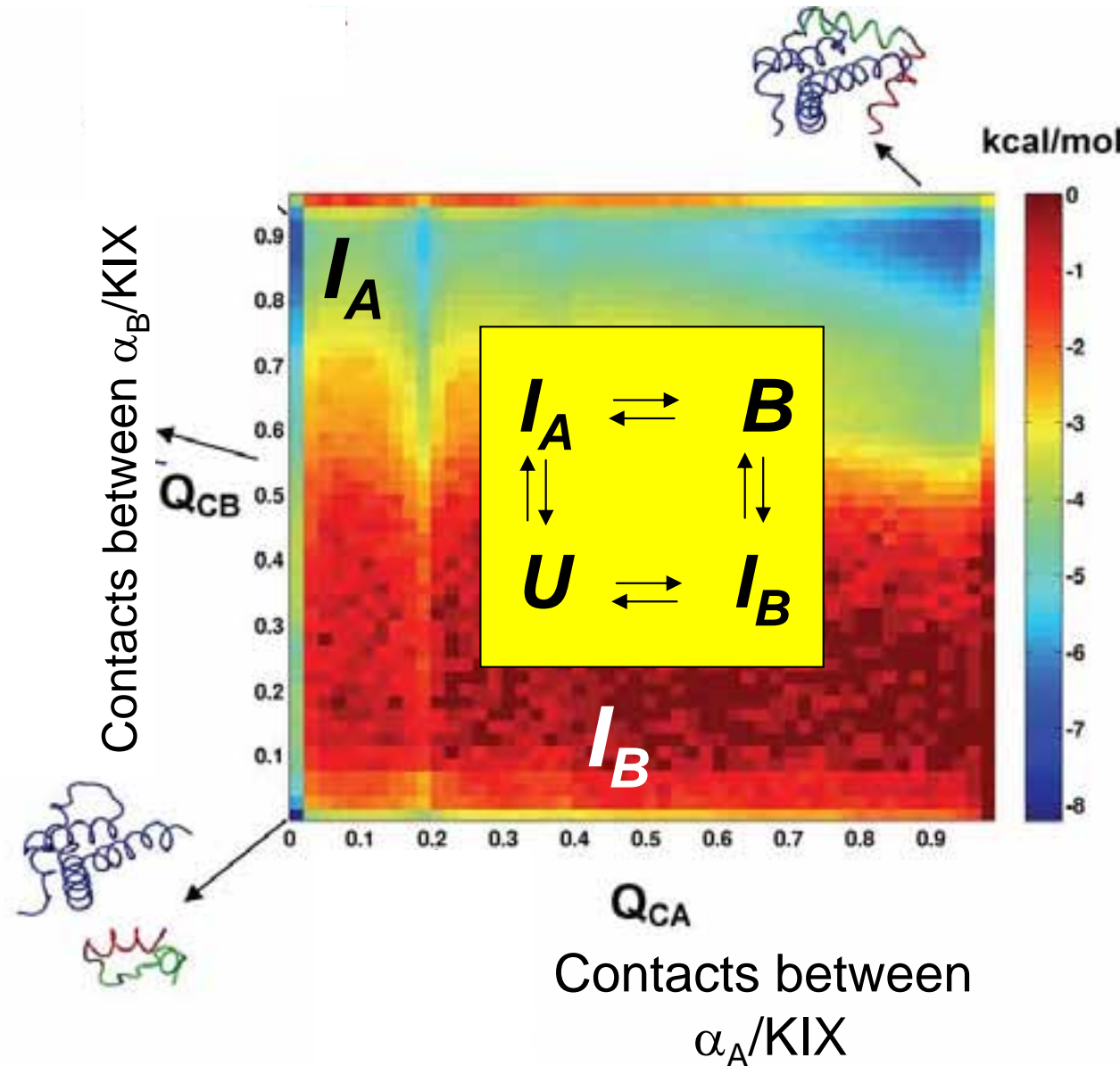
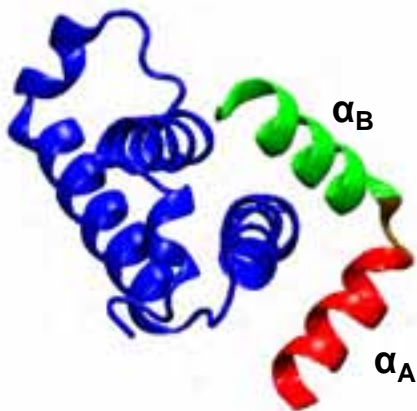
Fraction Q_B of α_B contacts

Binding proceeds through intermediates

(Turjanski, Gutkind, Best, GH, *PLoS Comp. Biol.* 4, e1000060, 2008)

- Simulations show two intermediates

- I_A has helix A dissociated and fully unfolded
- I_B has helix B dissociated and partially unfolded



Validation by NMR experiments

(Turjanski, Gutkind, Best, GH, *PLoS Comp. Biol.* **4**, e1000060, 2008)

- “Folding-after-binding” broadly consistent with conclusions from NMR experiments (Sugase, Dyson, Wright, *Nature* **447**, 1021, 2007)
- Presence of two intermediates is consistent with NMR data

Cluster	Residue	I→B	B→I
		k_{IB} (s ⁻¹)	k_{BI} (s ⁻¹)
α_A	I Arg 124	394 (8) _‡	928 (10) _‡
	I Arg 125	Slow	Fast
	I Ile 127	$k_{BI}/k_{IB} \sim 2.4$	
	I Leu 128		

Cluster	Residue	I→B	B→I
		k_{IB} (s ⁻¹)	k_{BI} (s ⁻¹)
α_B	II pSer 133	1718 (10)	224 (10)
	II Tyr 134	Fast	Slow
	II Arg 135	$k_{BI}/k_{IB} \sim 0.13$	
	II Lys 136		
	II Leu 138		
	II Asn 139		

Table 1 from Sugase, Dyson, Wright, *Nature* **447**, 1021, 2007

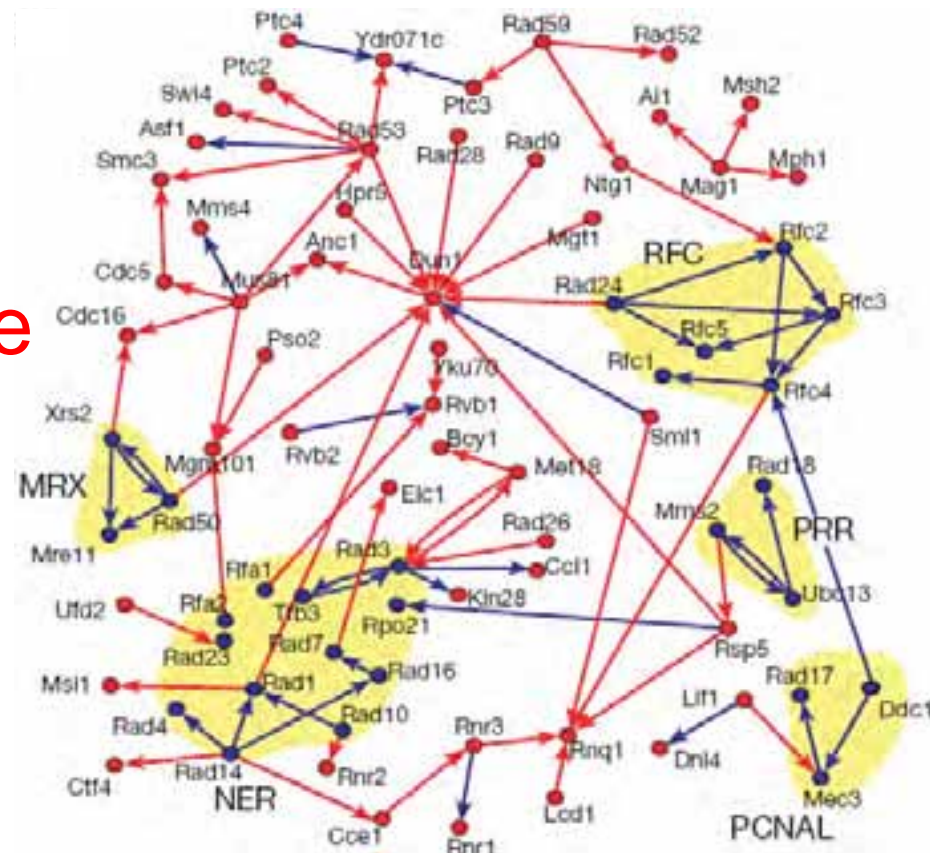
- Relative population from NMR kinetics is $[I_A]/[I_B] \sim 2.4/0.13 \sim 20$
- Relative population in simulations is $[I_A]/[I_B] \sim 70$
- “Error” $\sim kT \ln(70/20) \sim 0.75$ kcal/mol
- Structure agrees with NMR
 - α_A nearly fully folded in intermediate
 - α_B incompletely folded in intermediate

Conclusions (II): Transcription-factor binding

- Simple, topology-based coarse graining produces results broadly consistent with experiment
 - Binding of natively unstructured pKID/CREB to KIX/CBP precedes folding
 - Simulations suggest the presence of two intermediates, with either the A helix or the B helix dissociated
- Possible advantages of being unstructured?
 - No reduction of kinetic on-rate of binding
 - We find that unstructured peptide binds faster than pre-structured peptide, consistent with “fly-casting mechanism” of Shoemaker, Portman, Wolynes, *Proc Natl Acad Sci USA* **97**, 8868, 2000)
 - Ability to bind multiple targets
 - Rapid termination of the signal via peptide degradation

3. Spatial coarse graining: transferable energy functions for dynamic multi-protein assemblies

- **Proteomics**
 - Protein interaction networks dictate cellular function
- **Many biological functions are carried out by large, multi-protein assemblies**
 - DNA transcriptional regulation
 - Signal transduction
 - Nuclear pore complex
 - Membrane-protein trafficking
 - Viral assembly, entry and release



Yeast DNA damage response network.
Ho et al., *Nature* **415**, 180, 2002.

Motivation

- Many multi-protein assemblies form only transiently
 - Held together by relatively weak pairwise interactions ($K_d > 1\mu\text{M}$)
- Entire regions are often “natively unstructured”
 - Unfolded in the dissociated state
 - Structured in the bound state
- ⇒ Challenges for traditional structural-biology approaches
 - Crystallization of weak complexes with unstructured regions
 - NMR size limits
 - Trapping of functional assemblies for electron microscopy
- ⇒ Challenges for traditional modeling/simulation approaches
 - System size
 - Time scales of transient binding/dissociation and folding/unfolding
- ⇒ New opportunities for multi-scale modeling & simulation

Challenges

- **Competing protein-protein interactions**
 - *Binding affinities* of protein complexes need to be reproduced, in addition to their structure
- **Dynamic protein-protein interactions**
 - Create *equilibrium ensembles of bound structures*, including “transient encounter complexes,” not just single docked structures
- **Assumption**
 - In weakly-bound complexes, detailed atomic interactions (e.g., specific hydrogen bonds) are less important than interface complementarity (shape, charge, hydrophobicity)

Model and method

(Kim, GH, *J. Mol. Biol.* **375**, 1416, 2008)

- **Residue-level coarse-graining**

- Rigid body for folded domains

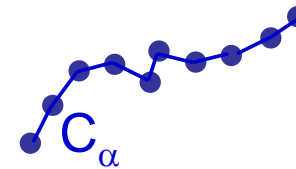


- **Transferable energy function**

- Long-range Debye-Hückel electrostatic interactions
- Residue-dependent short-range interactions (based on statistical contact potentials of Miyazawa and Jernigan, *J. Mol. Biol.* **256**, 623, 1996)
- Relative strength from *osmotic protein second-virial coefficients* and ubiquitin-UIM binding affinity (no structure!)

- **Flexible linkers: polymer model from protein folding studies**

- Harmonic stretching potential
- Bending potential
- Torsion angle potential

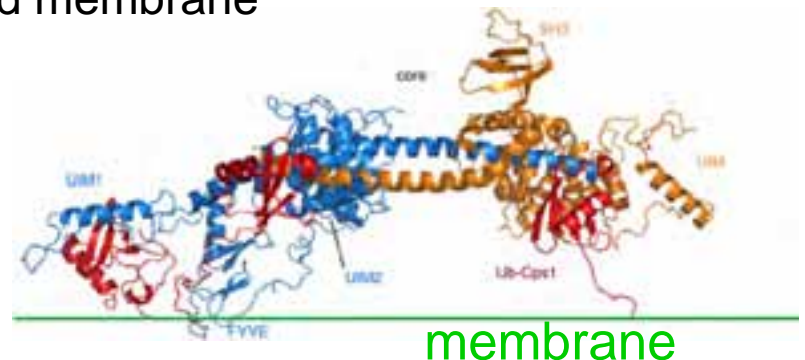


- **Membrane interactions**

- Planar membrane
- Statistical potentials between residues and membrane
- Electrostatic interactions

- **Replica exchange Monte Carlo**

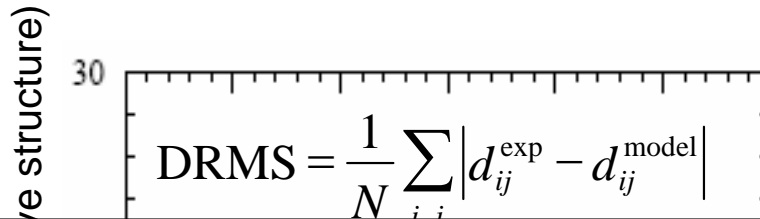
- Enhances equilibrium sampling



Validation: complex structure

(Kim, GH, *J. Mol. Biol.* **375**, 1416, 2008)

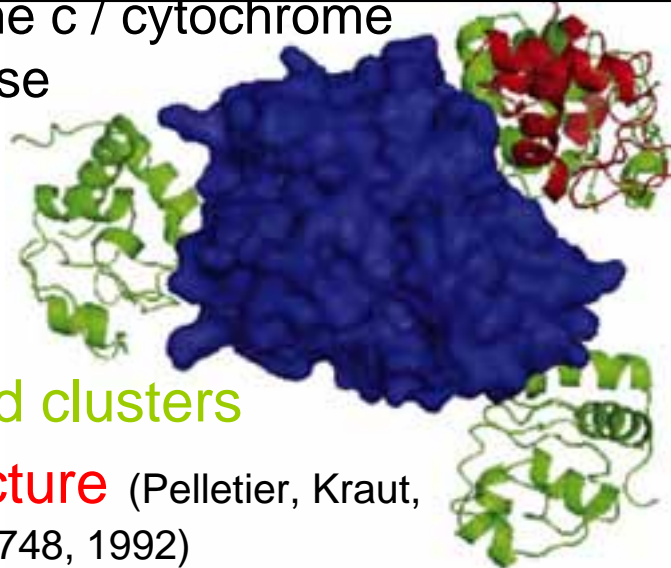
Ubiquitin/UIM1
structure prediction



Test set of >20 complexes:

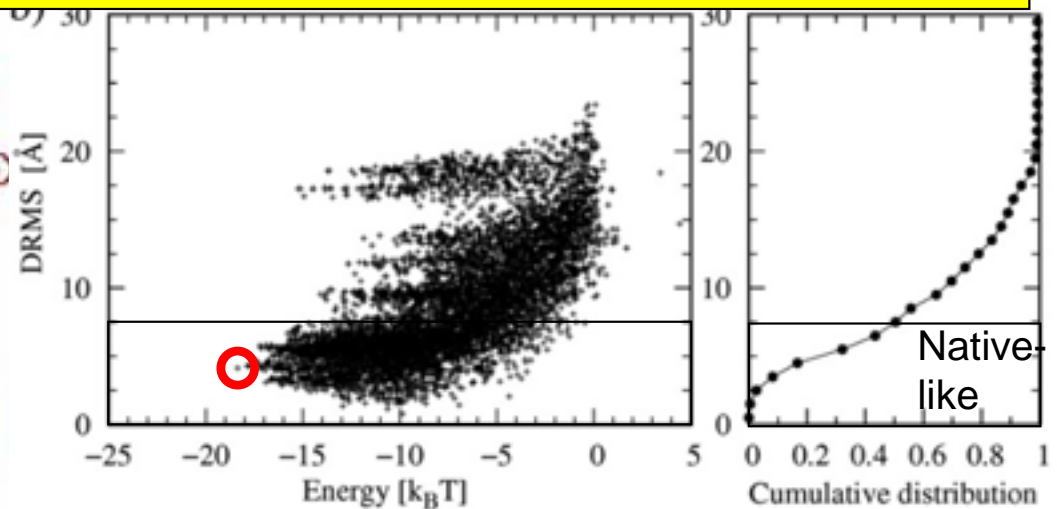
- In ~70% of the complexes, most populated cluster is within 5 Å dRMS
- In ~90% of the complexes, the binding interfaces on both proteins predicted correctly
- In all other cases, at least one interface is correct

Cytochrome c / cytochrome c peroxidase



Top-ranked clusters

X-ray structure (Pelletier, Kraut, *Science* **258**, 1748, 1992)

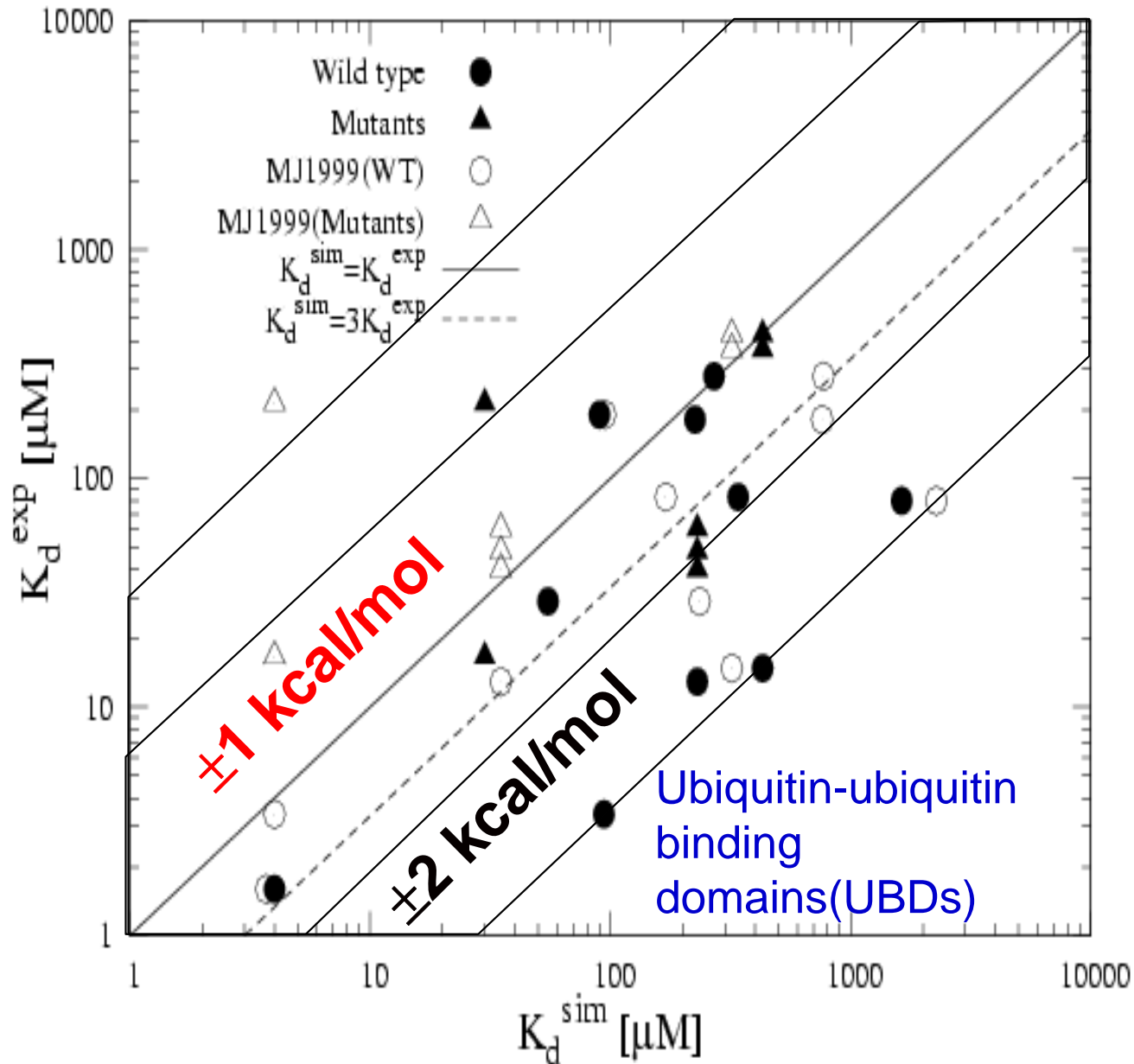


University, 5 Oct 2009

Validation: binding affinities

(Kim, GH, *J. Mol. Biol.* 375, 1416, 2008)

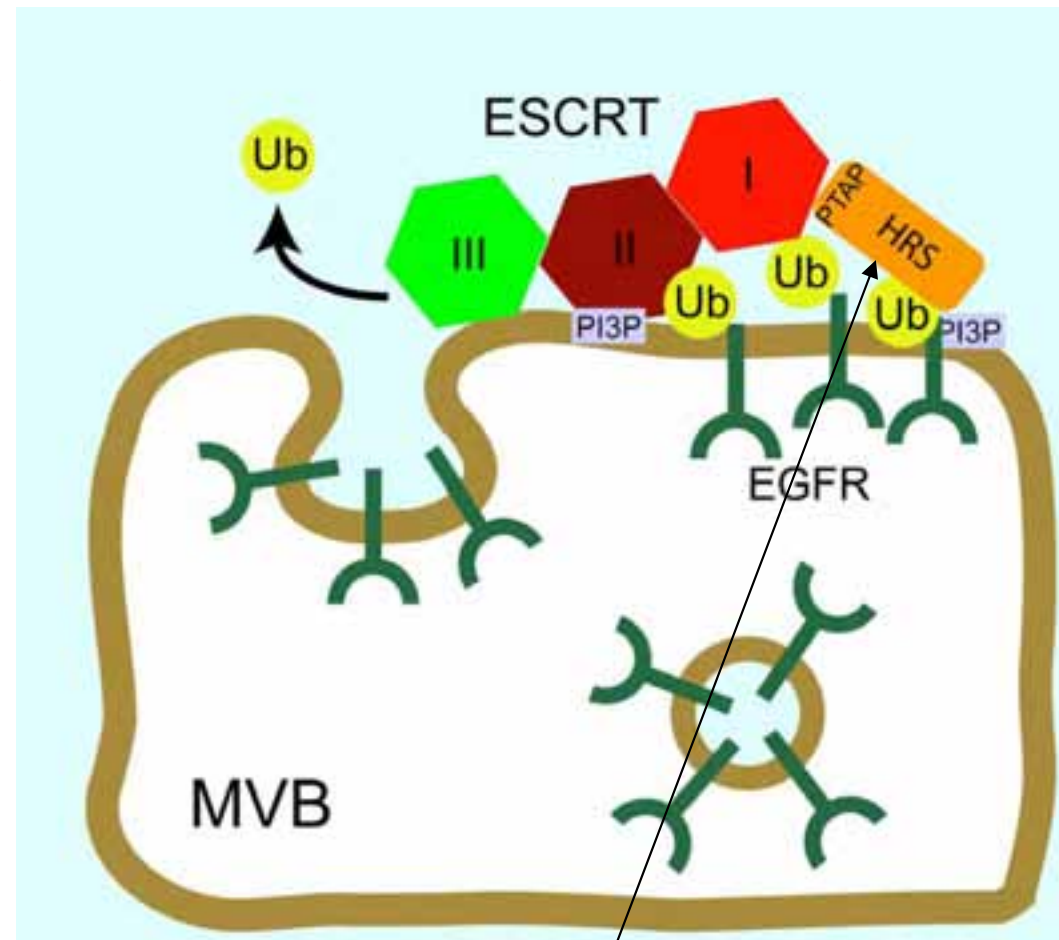
Ub/CUE
Ub/UIM1
Ub/UBA
Ub/GAT
Ub/DUIM
Ub/UIM2



Application: multi-vesicular body (MVB) protein sorting machinery

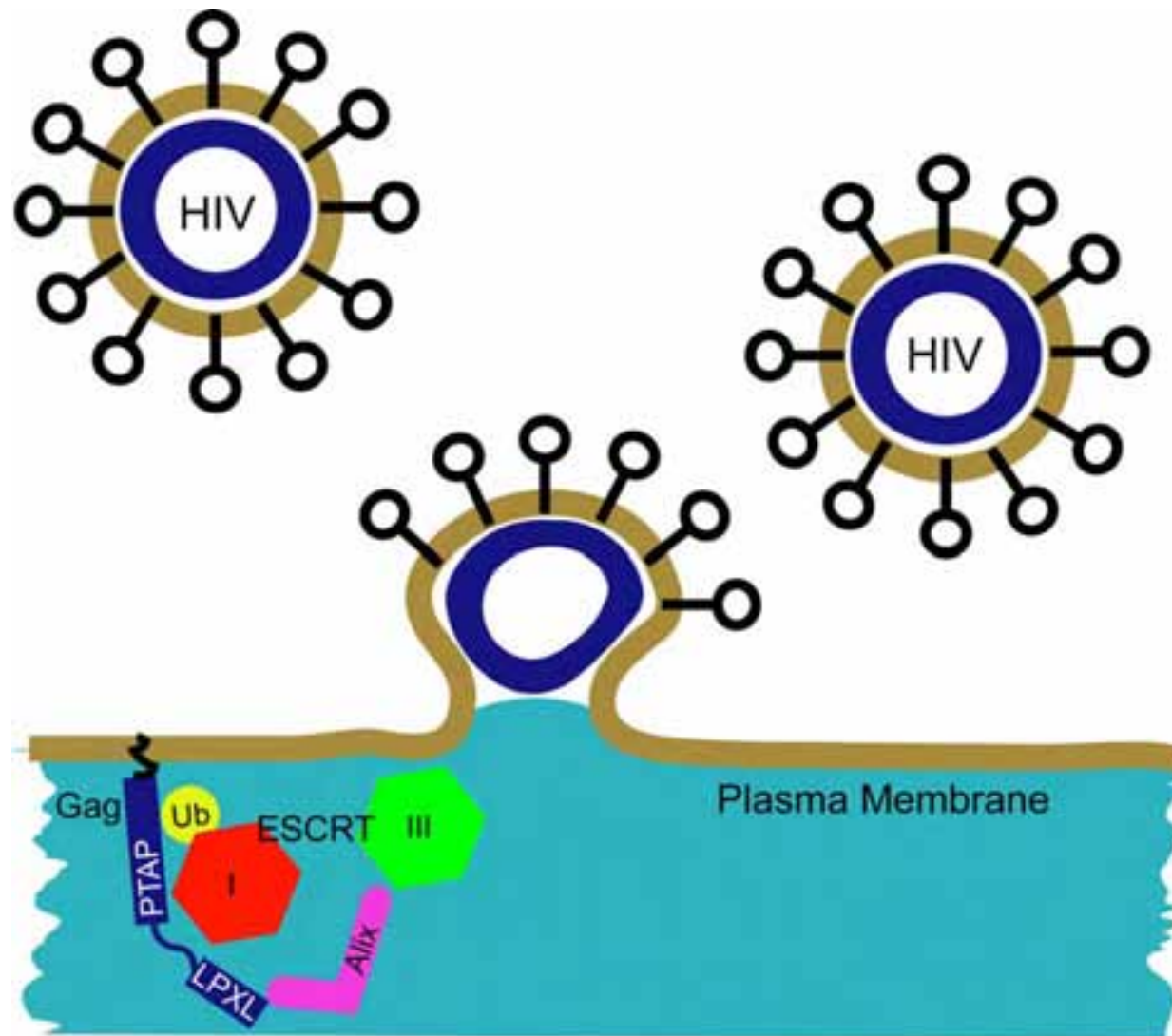
(Prag, Watson, Kim, Beach, Ghirlando, GH, Bonifacino, Hurley, *Dev. Cell* 12, 973, 2007)

- The ESCRT protein network targets ubiquitinated transmembrane proteins for degradation in the lysosome or yeast vacuole
- Transmembrane proteins are sorted into small vesicles that bud into the lumen of endosomes, thus forming multivesicular bodies (MVBs)
- Important, for instance, for receptor down-regulation



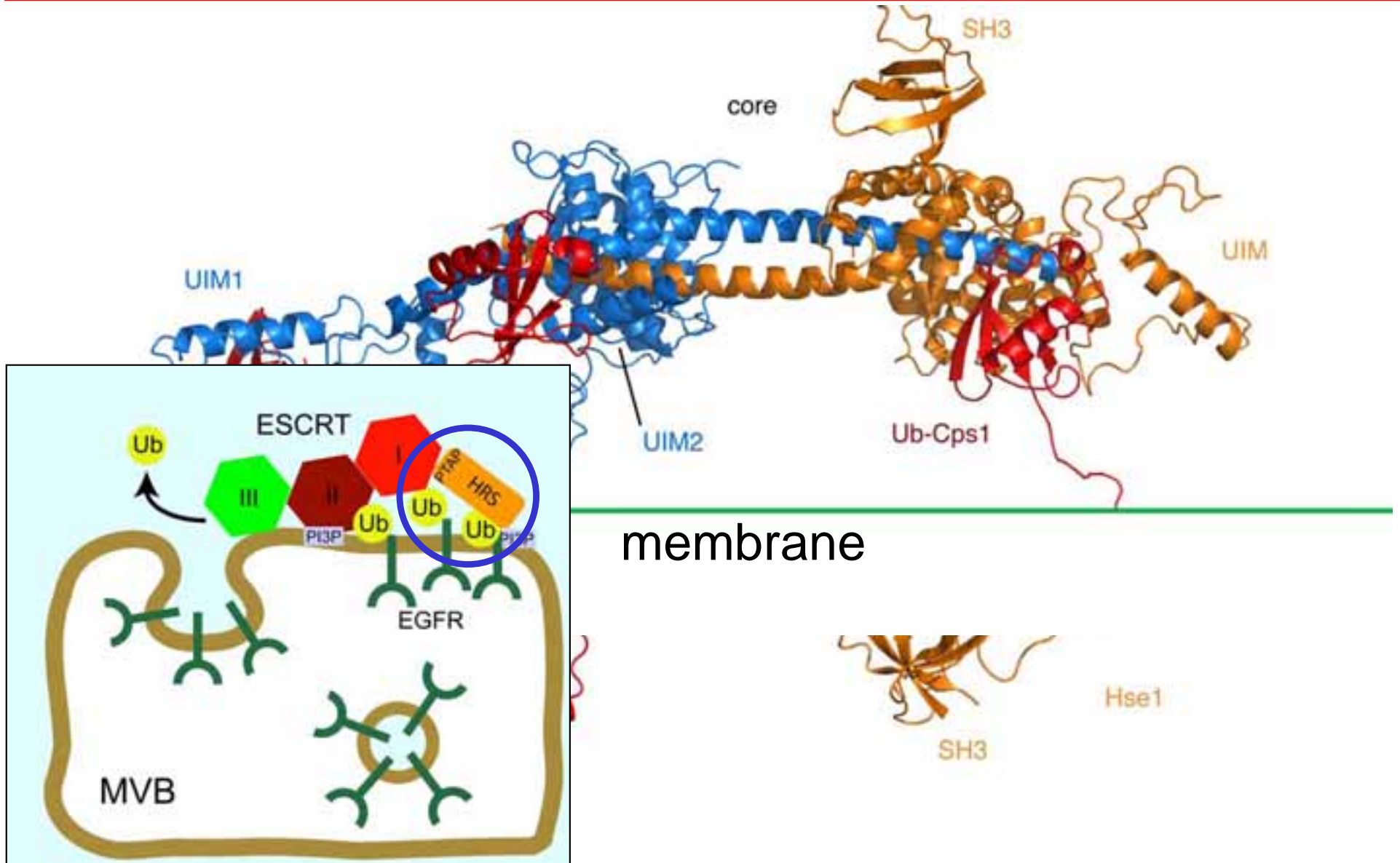
Vps27/Hse1 (yeast)
Hrs/STAM (human)

ESCRTs are required for HIV budding at the plasma membrane



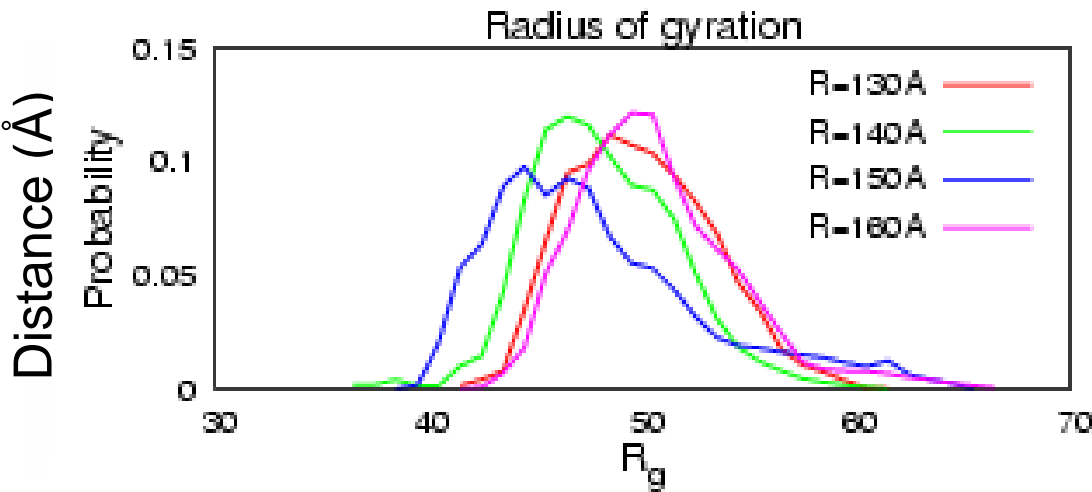
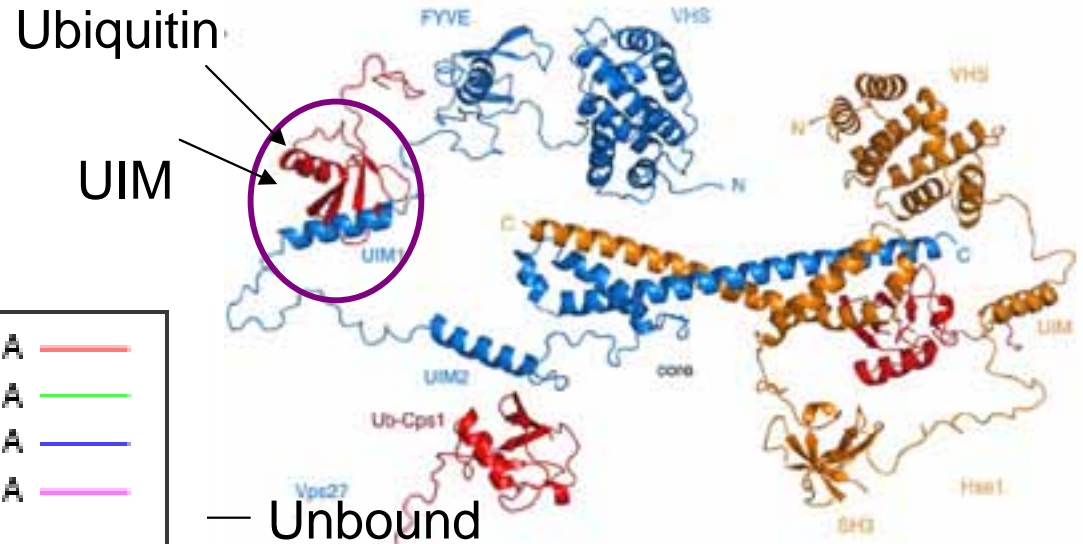
Structure of the assembled Vps27 complex

(Prag, Watson, Kim, Beach, Ghirlando, GH, Bonifacino, Hurley, *Dev. Cell* 12, 973, 2007)

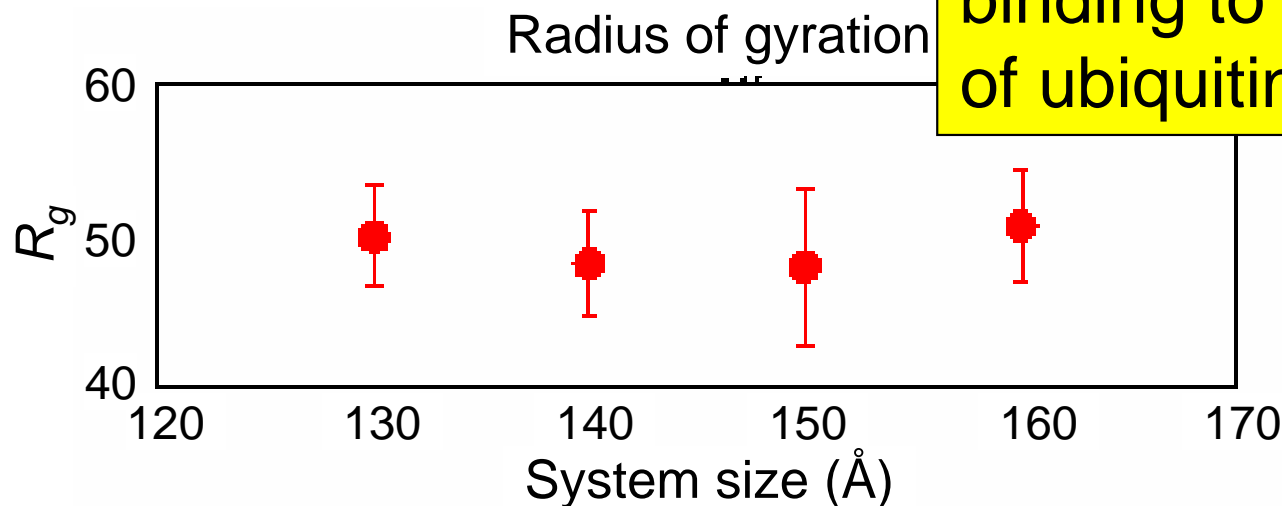


Vps27/Hse1 complex is dynamic and open

(Prag, Watson, Kim, Beach, Ghirlando, GH, Bonifacino, Hurley, *Dev. Cell* 12, 973, 2007)



Dynamic and open structure important for binding to a large variety of ubiquitinated targets



Experimental validation of open structure: Hydrodynamic measurements of Hrs/STAM complex

(Ren, Kloer, Kim, Ghirlando, Saidi, GH, Hurley, *Structure* **17**, 406, 2009)

- Measurements of hydrodynamic radius of full-length Hrs₁₋₇₇₇/STAM₁₋₅₄₀ complex (human ESCRT0):

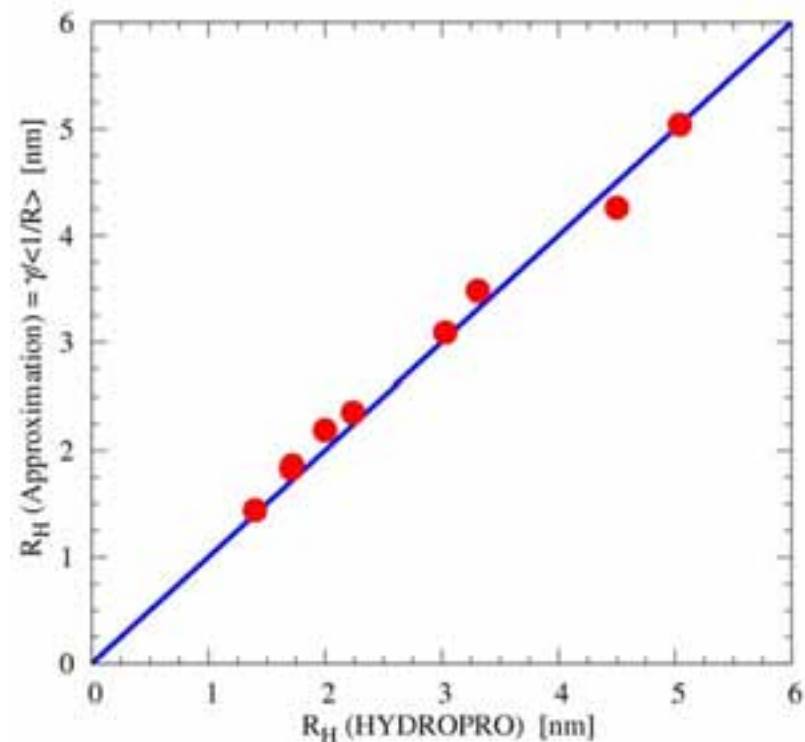
	SEC	AUC-DLS	QELS
R_H [nm]	7.8±0.10	8.0±0.05	7.9±0.10

- Simulation:**

- Combination of structured and large unstructured regions
- Polymer theory:

$$R_H = \gamma N^2 / \left\langle \sum_{i \neq j} 1 / r_{ij} \right\rangle$$

- $\gamma \approx 1.5$ fit to hydrodynamic calculations
- $\Rightarrow R_H \approx 6.8 \pm 0.6$ nm



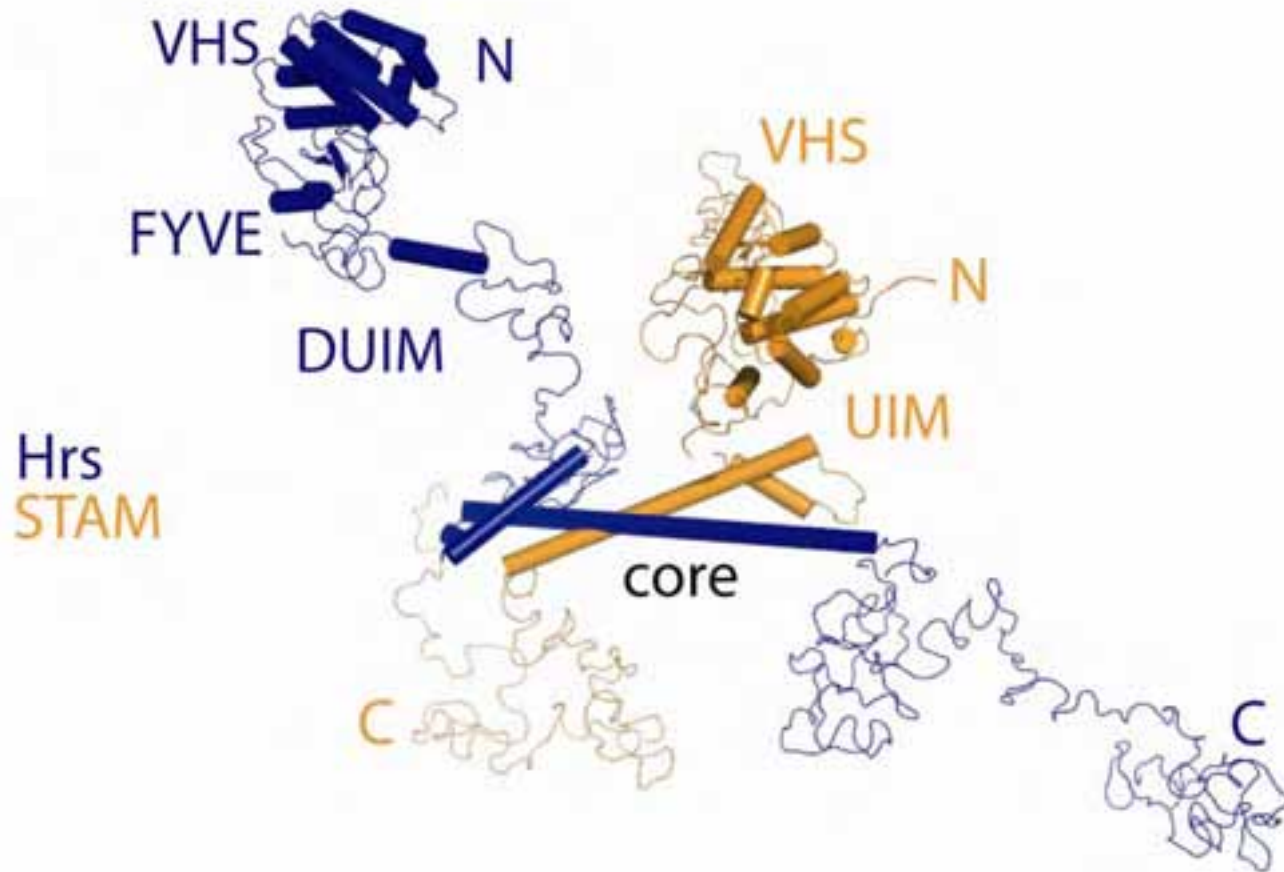
(HYDROPRO: Garcia de la Torre, Huertas, Carrasco, *Biophys. J.* **78**, 719, 2000)



Experimental validation of open structure: Hydrodynamic measurements of Hrs/STAM complex

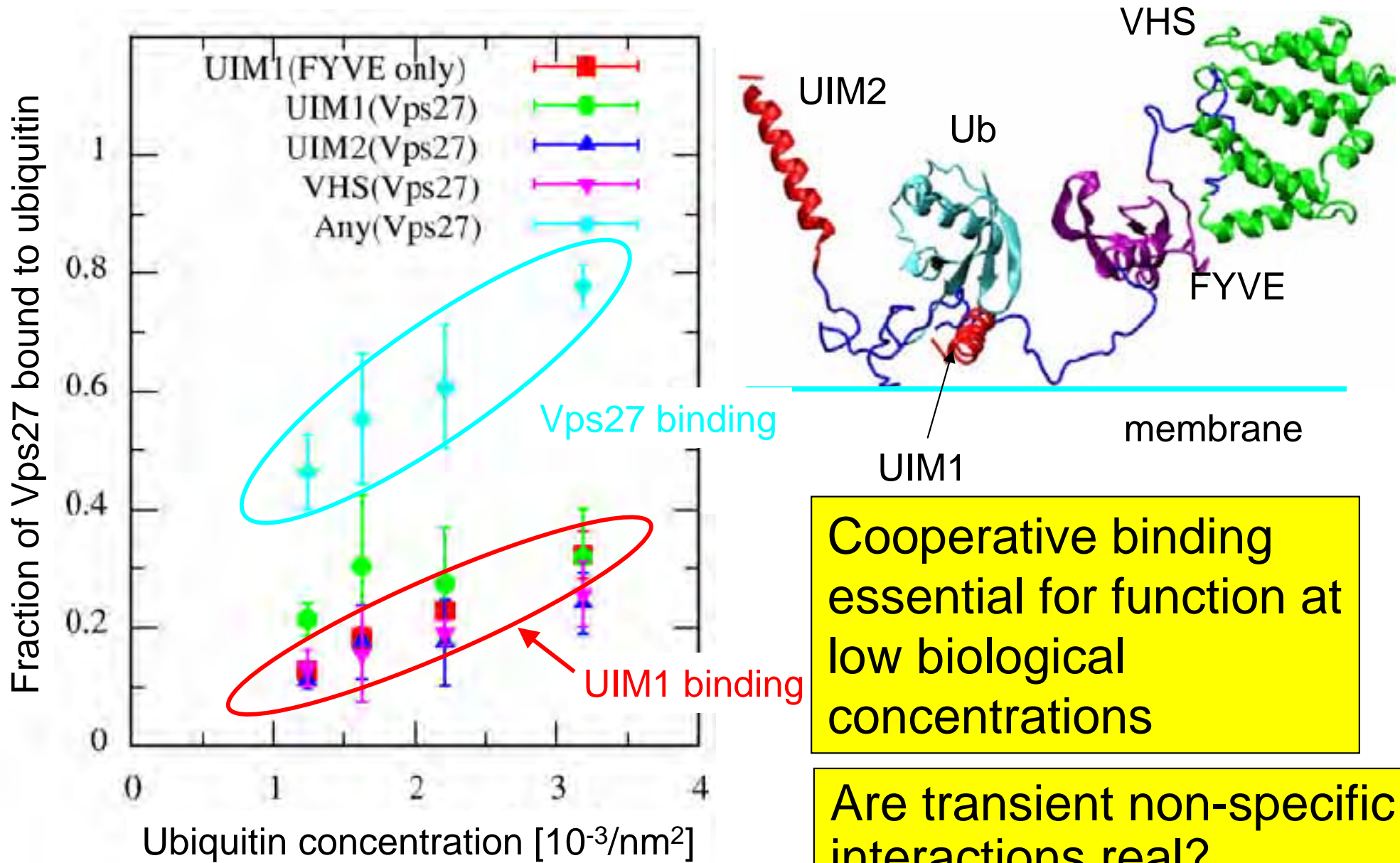
(Ren, Kloer, Kim, Ghirlando, Saidi, GH, Hurley, *Structure*, 17, 406, 2009)

- Snapshot of simulation restrained to experimental R_H



Positive cooperativity enhances Vps27 binding to ubiquitin

(Prag, Watson, Kim, Beach, Ghirlando, GH, Bonifacino, Hurley, *Dev. Cell* **12**, 973, 2007;
Kim, GH; *J. Mol. Biol.* **375**, 1416, 2008)



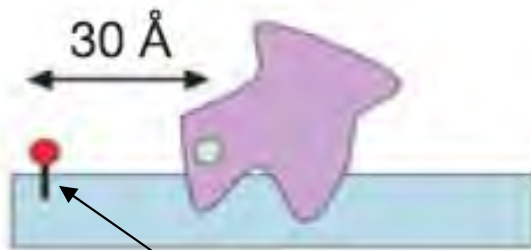
Transient protein encounter complexes probed by simulation and NMR

- Paramagnetic relaxation enhancement probes the presence of low-population (<10%) transient ($k_{ex} > 10 \text{ s}^{-1}$) encounter complexes

State A (major)

$$r = 30 \text{ \AA}$$

$$\Gamma_{2,A} = 2.0 \text{ s}^{-1}$$



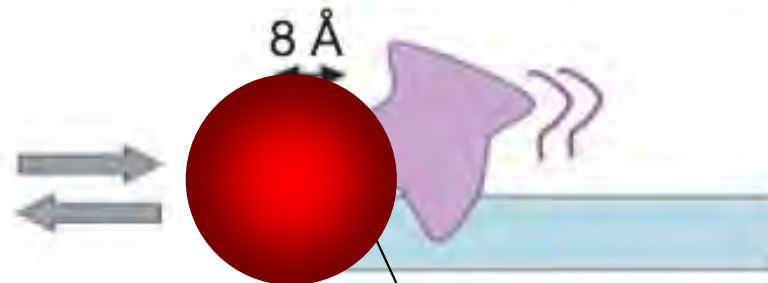
paramagnetic label
(unpaired electron)

High population
small PRE ($\sim r^6$, r large)

State B (minor)

$$r = 8 \text{ \AA}$$

$$\Gamma_{2,B} = 5.6 \times 10^3 \text{ s}^{-1}$$



proton spin probed

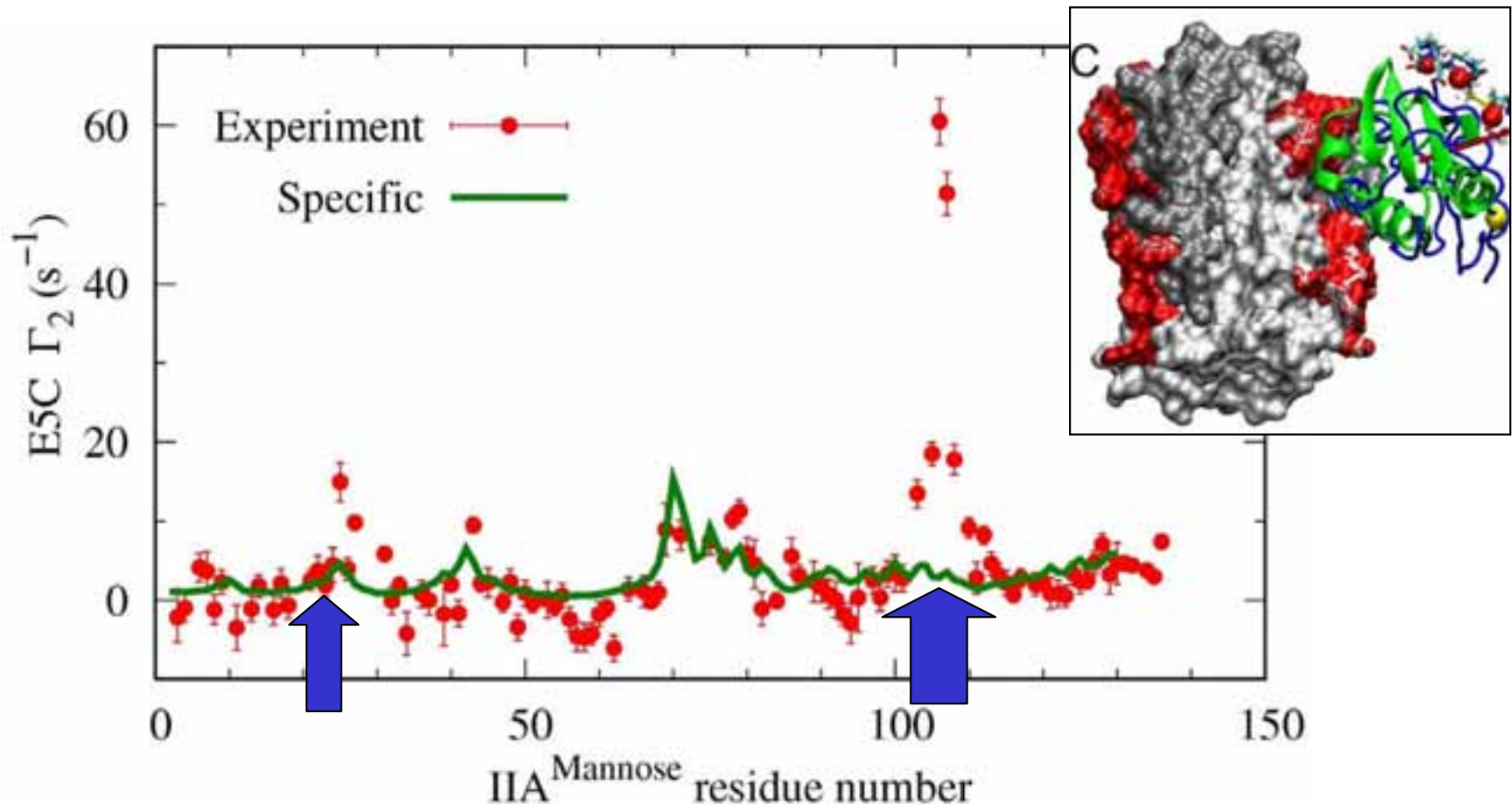
Low population
large PRE ($\sim r^6$, r small)

Iwahara, Clore,
Nature **440**,
1227, 2006

PRE profiles of HPr-IIA^{Man} complex of bacterial phosphotransferase system

(Tang, Iwahara, Clore, *Nature* **444**, 383, 2006)

- PRE of backbone amide protons on IIA^{Mannose}



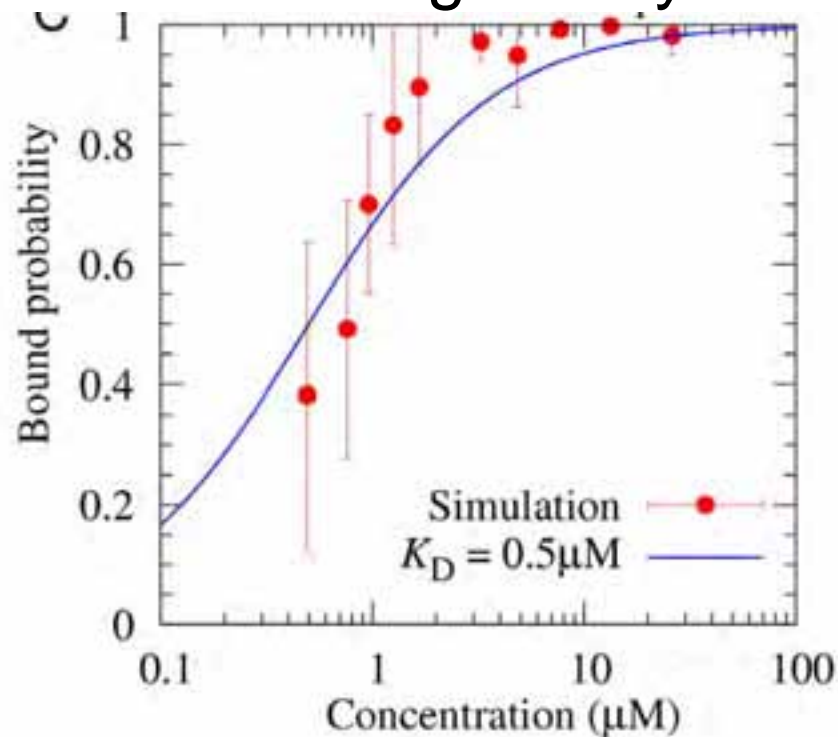
Can we simulate both the specific and non-specific binding?

Replica-exchange simulations of HPr-IIA^{Man} complex

(Kim, Tang, Clore, GH, *Proc. Natl. Acad. Sci. USA* **105**, 12855, 2008)

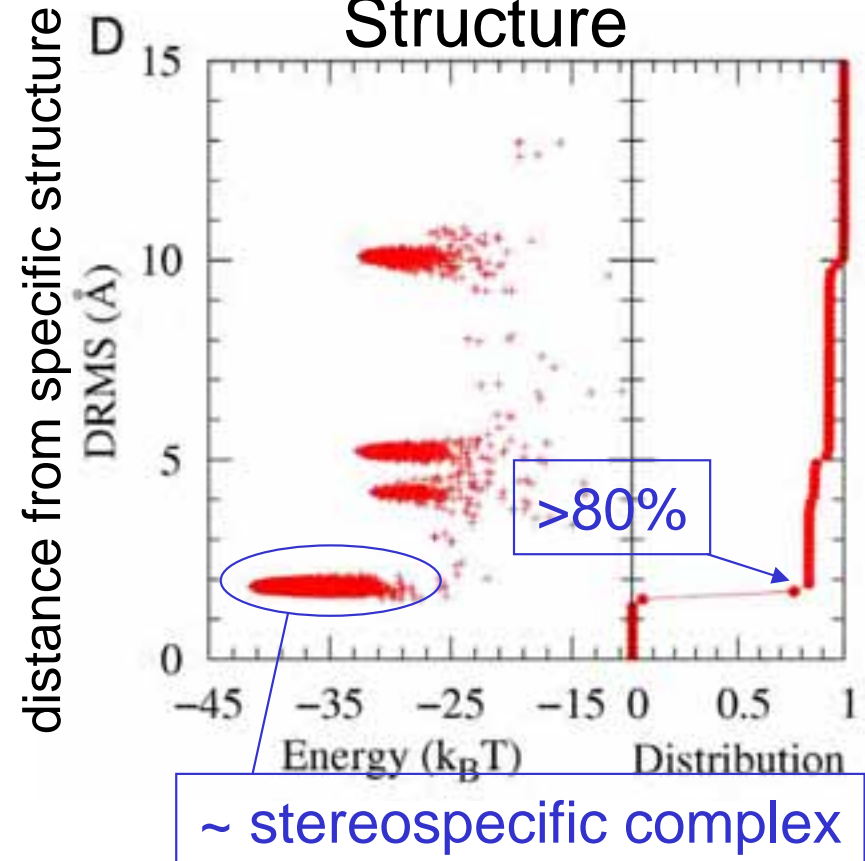
- **Coarse-grained simulation model** (Kim, GH; *J. Mol. Biol.* **375**, 1416, 2008)
 - adapted by sharpening of 12-6 contact potentials to 12-10-6 potentials (with a small barrier)

Binding affinity



$K_d(\text{exp}) \sim 30 \mu\text{M} \Rightarrow \text{error} \sim kT \ln 60 \sim 2.5 \text{ kcal/mol}$

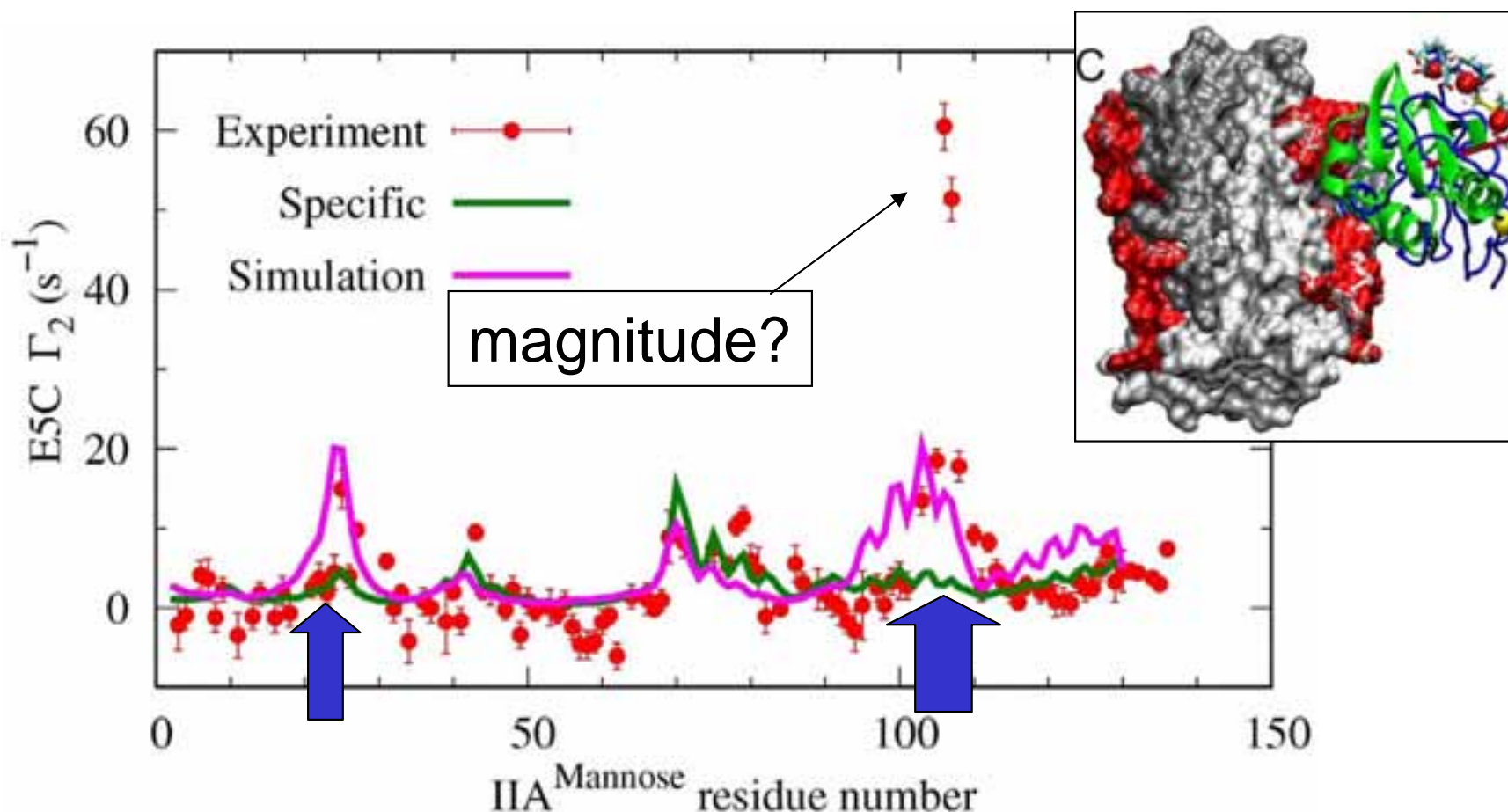
Structure



PRE profiles of HPr-IIA^{Man} complex

(Kim, Tang, Clore, GH, *Proc. Natl. Acad. Sci. USA* **105**, 12855, 2008)

- PRE of backbone amide protons on IIA^{Mannose}



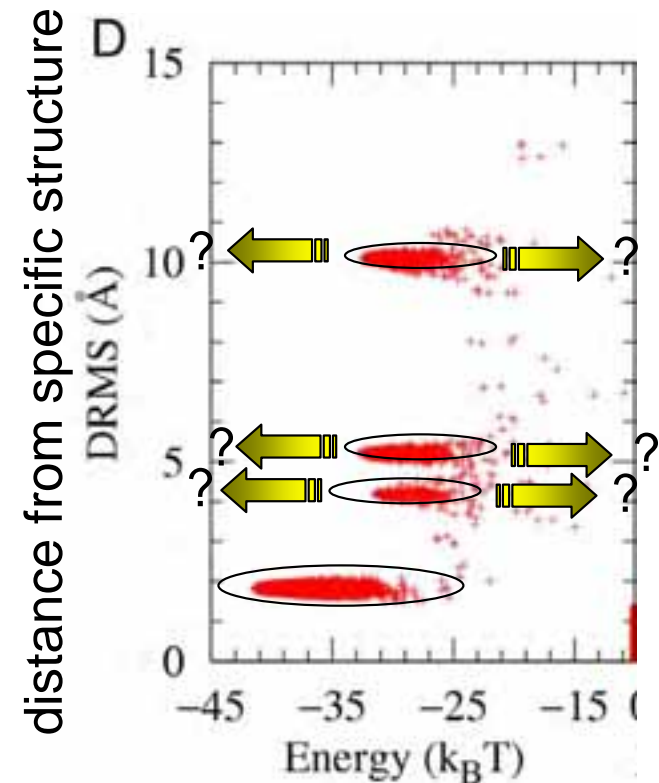
Reweighting of simulation structures

(Kim, Tang, Clore, GH, *Proc. Natl. Acad. Sci. USA* **105**, 12855, 2008)

- Simulation model should not be expected to be accurate to more than a few $k_B T$

⇒ Cluster the structures of the specific and non-specific complexes

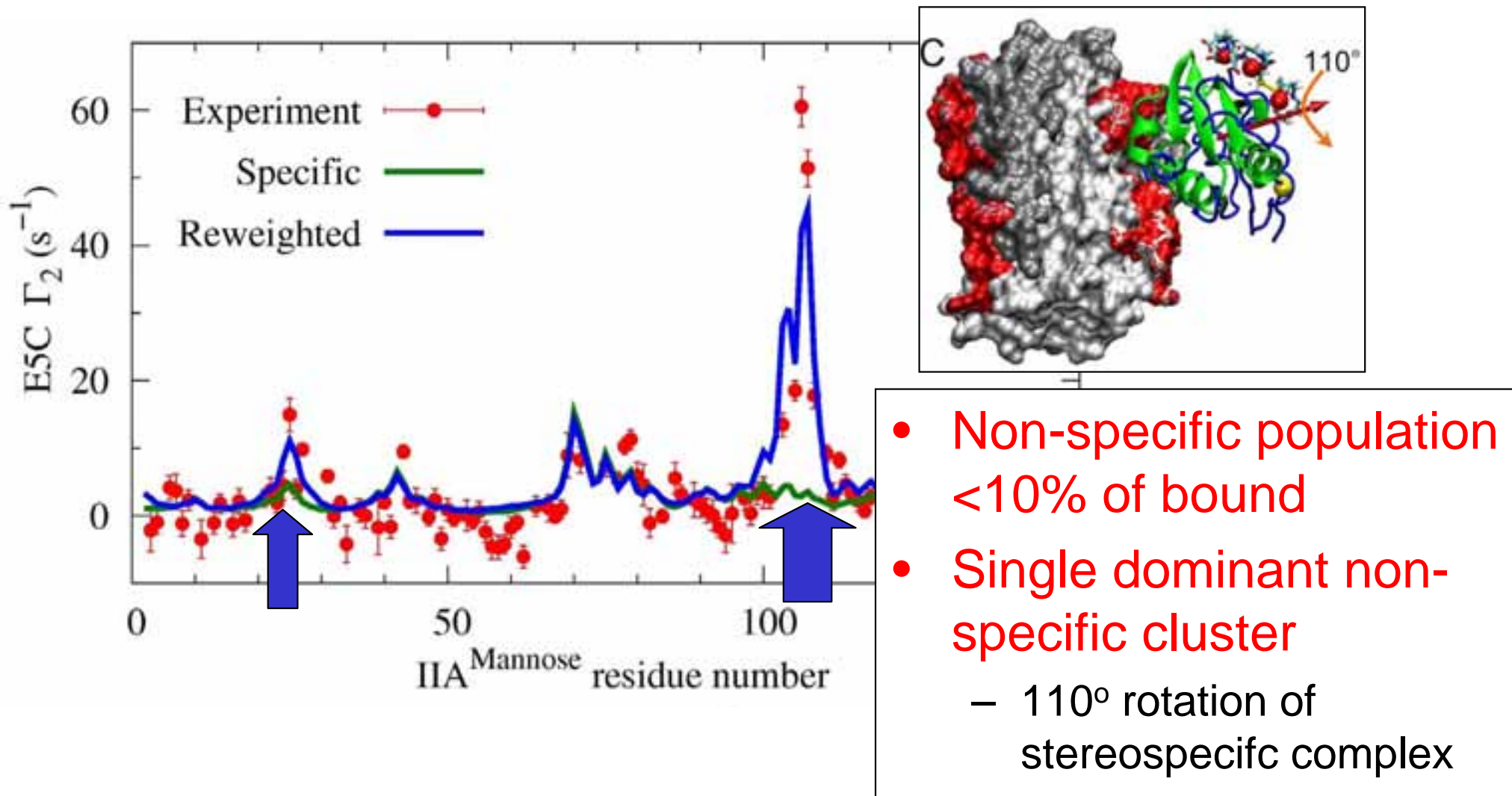
⇒ Re-weight the populations of the clusters to match PRE profiles



PRE profiles of HPr-IIA^{Man} complex

(Kim, Tang, Clore, GH, *Proc. Natl. Acad. Sci. USA* **105**, 12855, 2008)

- PRE of backbone amide protons on IIA^{Mannose}



Conclusions (III): Multi-protein assemblies

- Transferable energy function predicts structure and affinities of weak complexes
 - Why does such a simple model work so well? Overall dominance of hydrophobic interactions, decorated by charge for orientational specificity?
 - Easy incorporation low-resolution experimental data (R_H , SAXS, ...)
 - Fine-graining with molecular dynamics simulations
- ESCRT0 (Hrs/STAM in human, Vps27/Hse1 in yeast) forms open and dynamic complex to recognize ubiquitinated proteins
- Specific and non-specific interactions act cooperatively to function at low biological concentrations
- Transient non-specific complexes are an important element of the bound state

Outlook

- “Grand challenges” in biology, materials science, nanotechnology, ..., cover multiple scales
 - From atomistic to mesoscopic / macroscopic
 - From quantum chemistry of enzyme catalysis to the function of organelles and cells
- Molecular simulations are, arguably, the most powerful tool for theoretical studies of complex molecular systems
 - Simple and generally applicable
 - Rapidly growing demand on computer resources
- Non-equilibrium statistical mechanics provides theoretical framework to develop internally consistent multiscale approach

Acknowledgments

- **Master equation**
 - N.-Vio Buchete (NIH; University College Dublin)
 - Yannis Kevrekidis (Princeton)
- **Transcription-factor binding**
 - Adrian G. Turjanski (NIDCR, NIH; University of Buenos Aires)
 - Robert Best (NIH; University of Cambridge)
 - J. Silvio Gutkind (NIDCR, NIH)
- **Multi-protein complexes: Vps27/ESCRT**
 - Young C. Kim (NIH; NRL)
 - James Hurley (NIDDK, NIH)
- **Transient encounter complexes**
 - Young C. Kim (NIH; NRL)
 - G. Marius Clore (NIDDK, NIH)