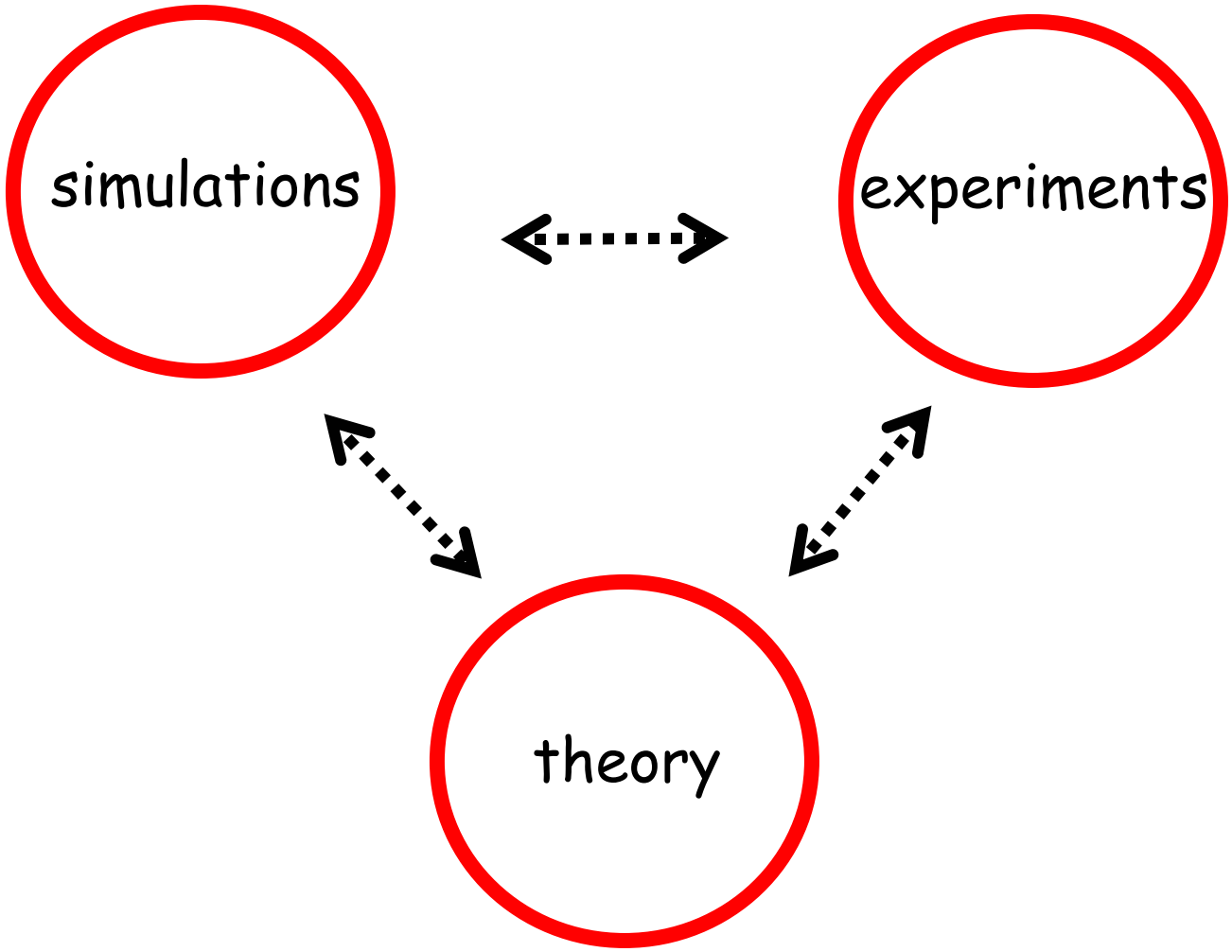
KITP
June 9, 2011

An Ising-like model for protein folding
(Kindergarten statistical mechanics)

William A. Eaton.
Laboratory of Chemical Physics
NIDDK, National Institutes of Health (NIH)
Bethesda, Maryland


simulations
experiments
theory

No completely realistic theoretical model that can quantitatively explain equilibrium and kinetic data on protein folding

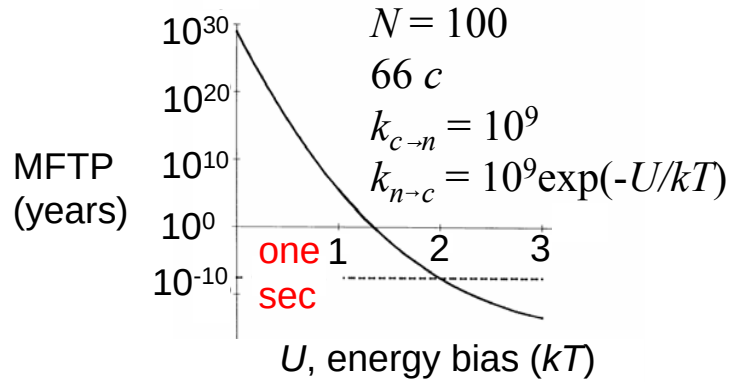
Theoretical work (and I don't mean MD simulations) has been introducing insightful concepts from modern statistical mechanics and simulations

My bias as experimentalist: theoretical model should consist of partition function, master equation, and theory for experimentally measured quantities

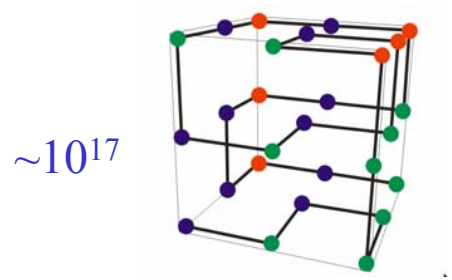
Muñoz/Henry/Eaton theoretical model motivated by 4 key results.

1. Debunking of Levinthal paradox by Zwanzig, Szabo, & Bagchi, *PNAS* 1992
Search 3^{100} conformations in 100 fs steps requires 10^{27} years. $2kT$ energy bias reduces search to 1 sec.

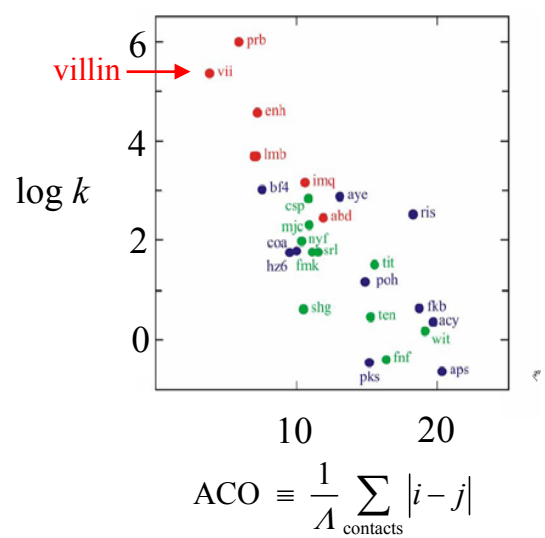
$$\ldots cccncncncccc \ldots \longrightarrow \ldots nnnnnnn \ldots$$



2. Demonstration by Onuchic and Wolynes in lattice simulations that diffusion on a one-dimensional free energy surface with order parameter as reaction coordinate can reproduce simulated folding rate (*JCP* 1996)

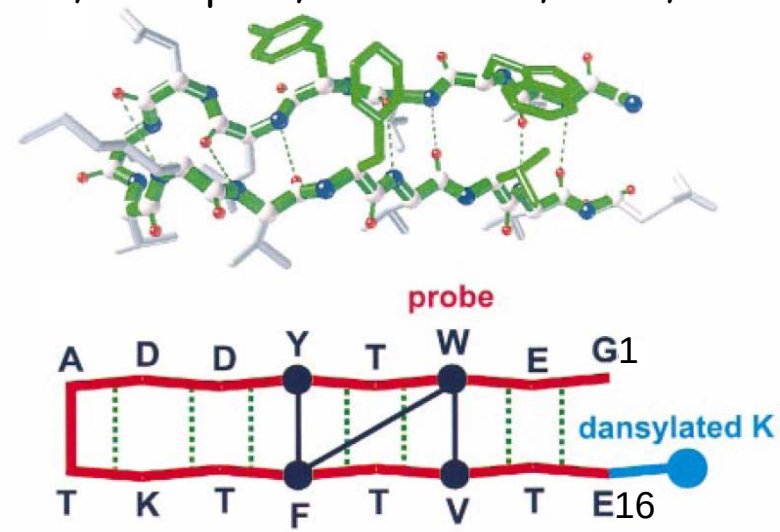


3. Observation of rate/contact order correlation by Plaxco and Baker – *JMB* 1998, knowledge of native structure is sufficient to predict rate; can ignore non-native interactions

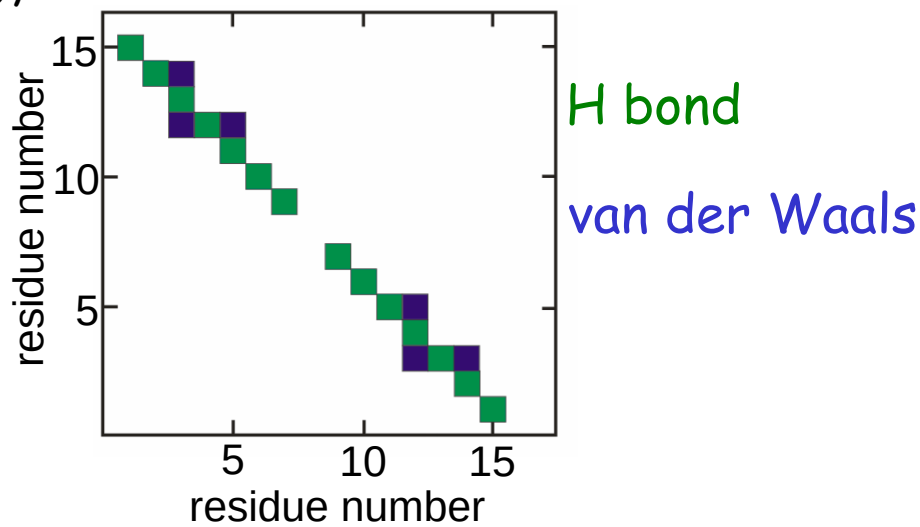


4. Success of Ising-like model

Muñoz, Thompson, Hofrichter, WAE, *Nature* 1997



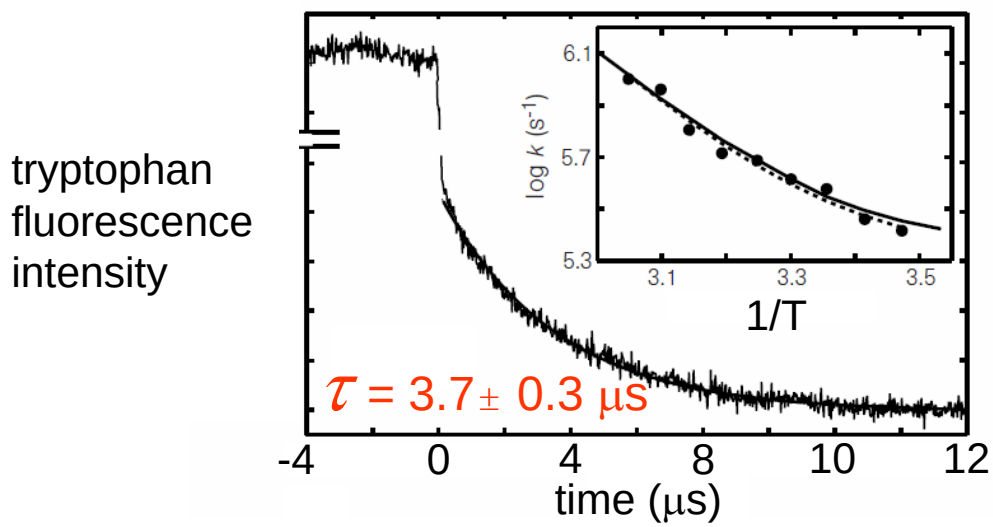
contact map of native structure



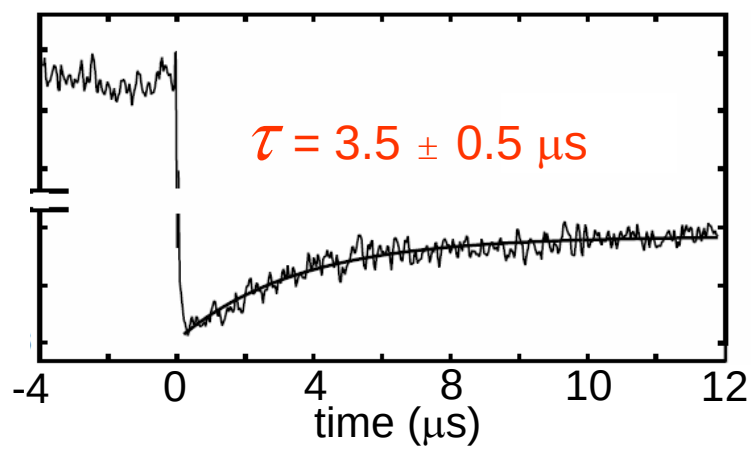
Two-state: $K_{eq} = k_f/k_u$; $k(\text{obs}) = k_f + k_u$

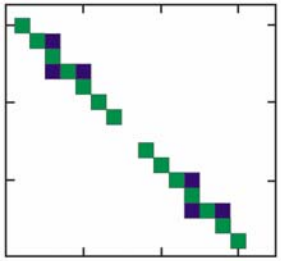
k_f slightly negative activation energy

tryptophan burial



FRET





Contacts only between residues in ***native*** conformation

Single sequence approximation (2^{16} = 65,536 → 121)

$$Z = \sum_{i=1}^{121} P_{eq,i} = \sum_{i=1}^{121} \exp\left(-\frac{Q_{HB,i}\varepsilon_{HB} + Q_{vdW,i}\varepsilon_{vdW} - N_i T \Delta s_{conf}}{k_B T} \right)$$

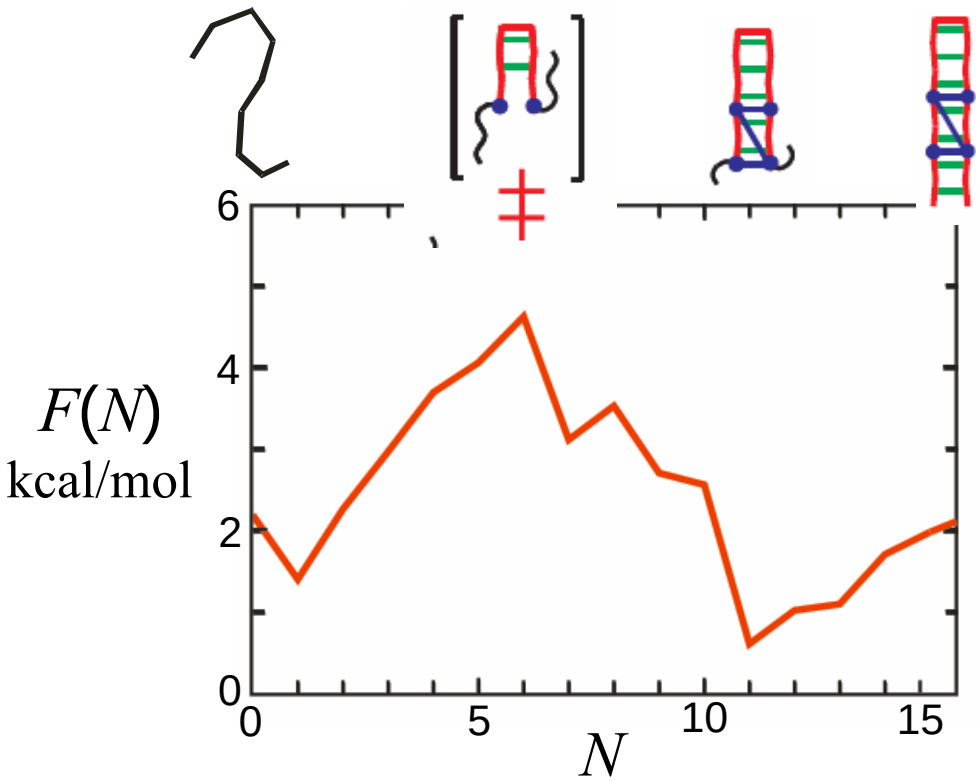
cccnnnnnccccccc

not

cccnncnnnccccccc

$$F(N) = -RT \ln\left(\sum_{\substack{\text{all configurations } i \\ \text{with } N \text{ native residues}}} P_{eq,i} \right)$$

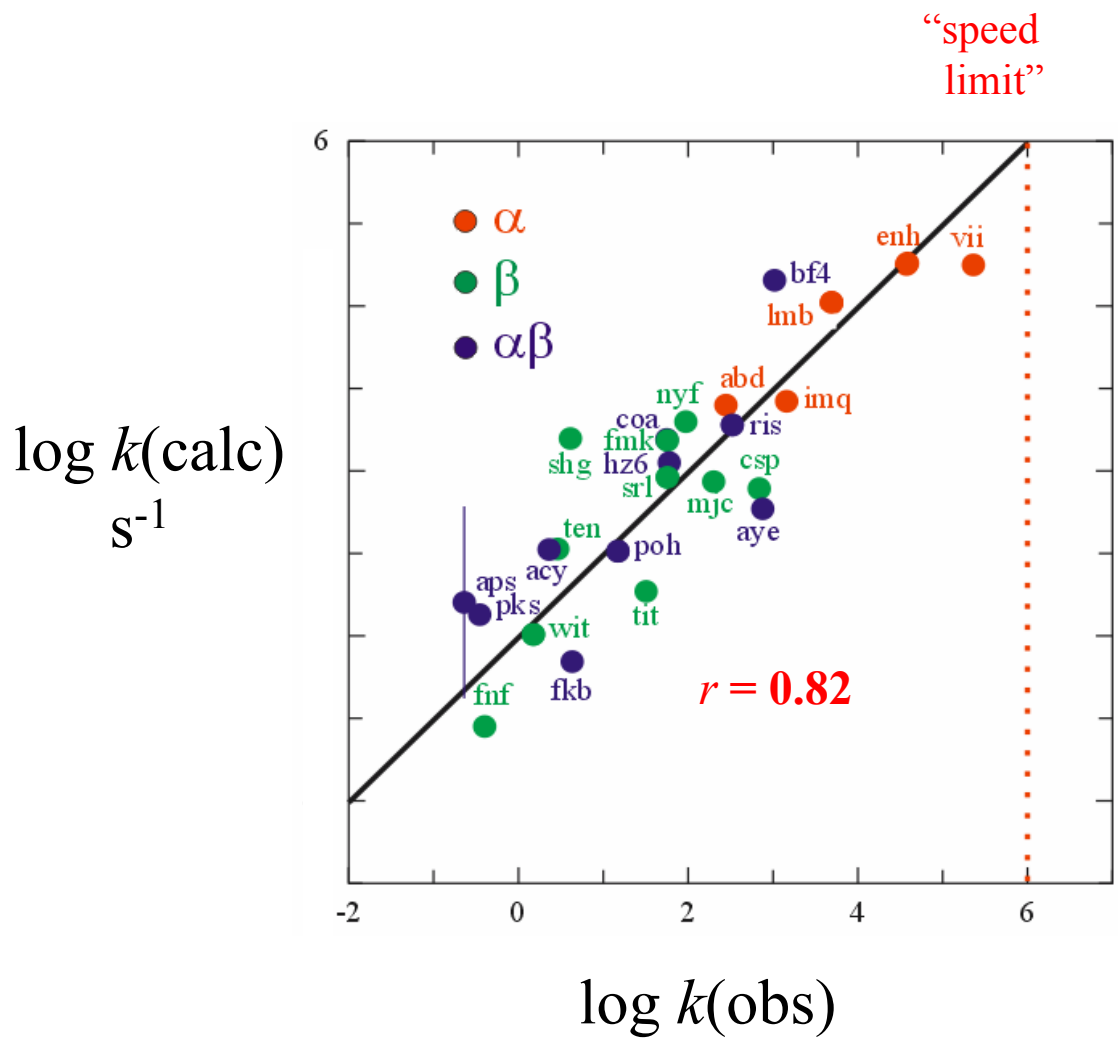
Correct equations
Video unclear



Ising-like model explained
2-state behavior
location of folded minimum
negative E_a for folding

Predicted mechanism:
now observed using most accurate force field
(Best & Mittal, PNAS,in press)

Same conformational entropy change and diffusion coefficient for all proteins – adjust contact energy to fit experimental equilibrium constant



ccccnnnnnnccc...
c
c ||
c
*cccc*nn*ccccccc*...

double sequence approximation

Next challenge

Explain equilibrium and kinetic observables, not just number of states and relative rates

Kubelka, Henry, Cellmer, Hofrichter and WAE, *PNAS* 2008

Why study ultrafast folders?

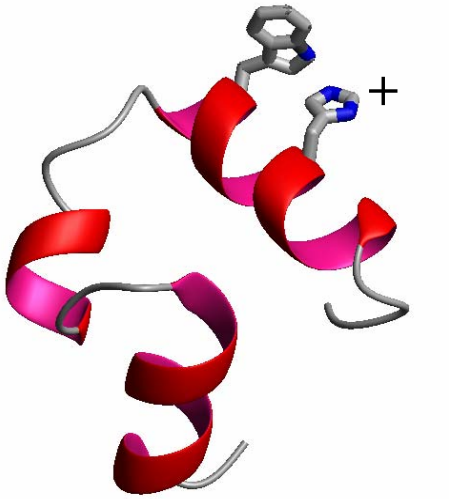
Can be simulated by all atom molecular dynamics calculations

Small proteins, and therefore simplest mechanisms

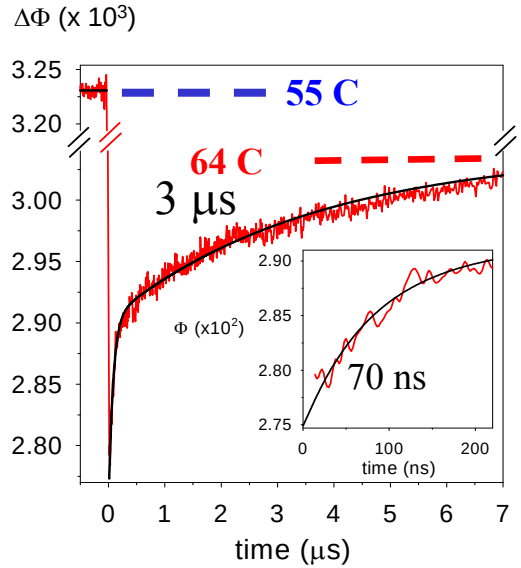
Therefore make connections among experiment, simulations, and theory

In 2003 we guessed (correctly) that the villin headpiece subdomain would be one of the most widely studied proteins by experiment, theory, and simulations
(98 papers from 45 different groups
(40 theoretical/simulation) as of December 2010).

Villin headpiece subdomain



quantum yield ($\times 10^2$)



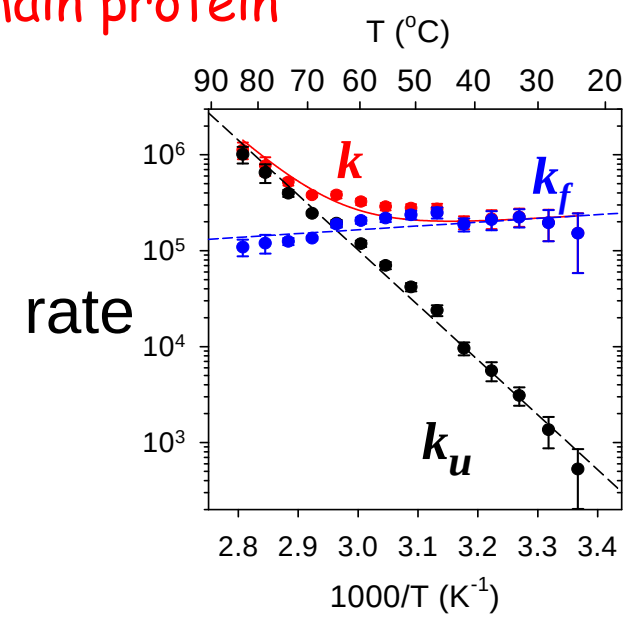
35 residues: smallest naturally occurring sequence that behaves like a typical single domain protein

Two-state analysis ($U \rightleftarrows F$):

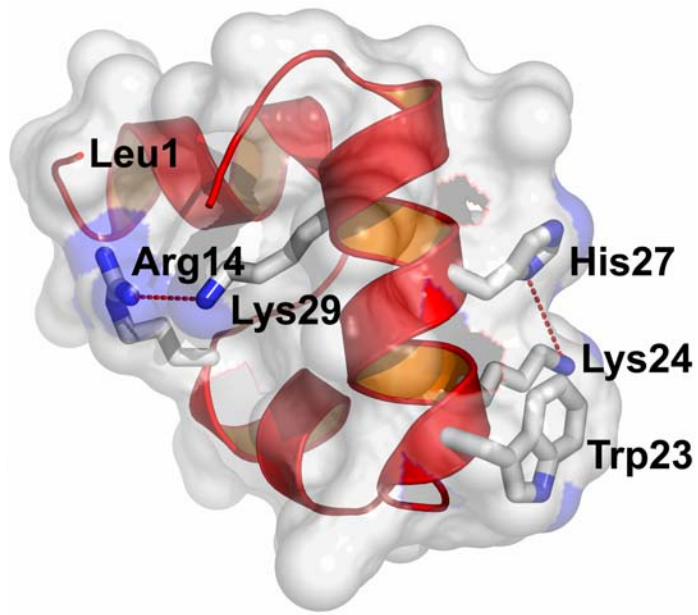
$$k(\text{obs})_{\text{obs}} = k_f + k_u$$

$$K_{\text{eq}} = k_f / k_u$$

$$\tau_{\text{folding}}(27^{\circ}\text{C}) = (4.3 \pm 0.6)\ \mu\text{s}$$



Short Digression

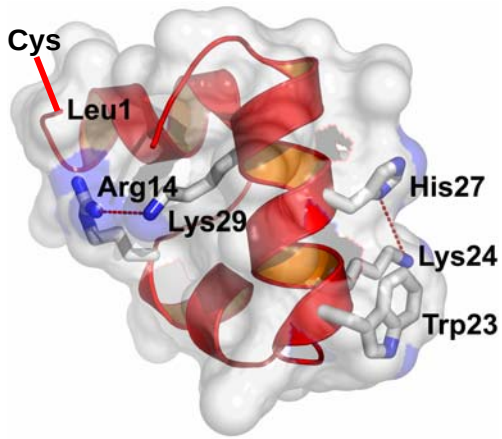


Ensign and Pande, J Mol Bio 2007

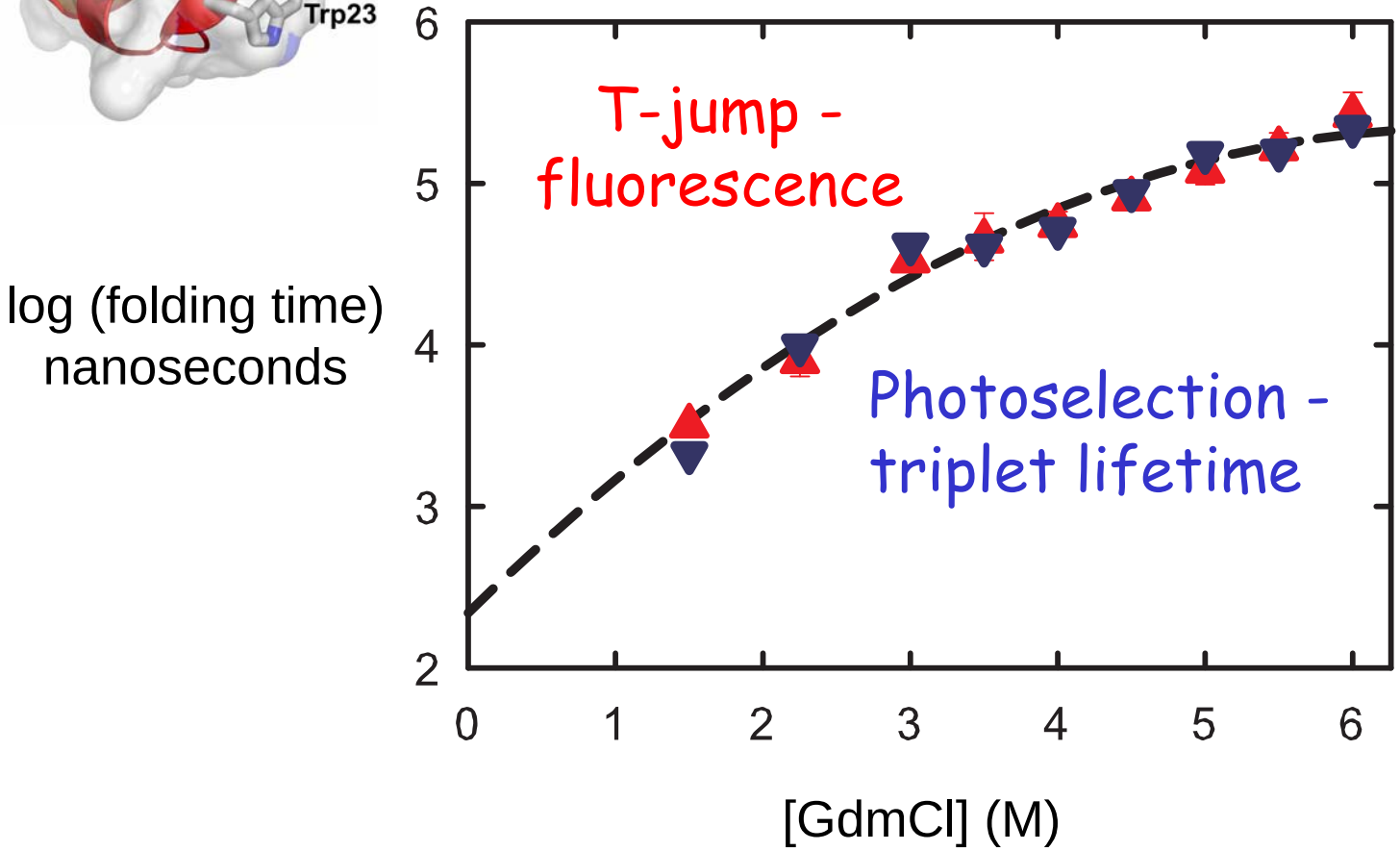
Freddolino and Schulten, Biophys J 2009

Fluorescence does not monitor folding

In Pande simulations slowest fluorescence change is up to 20 times faster than folding rate



Cellmer, Buscaglia, Henry, Hofrichter & WAE
Proc. Natl. Acad. Sci. USA 2011

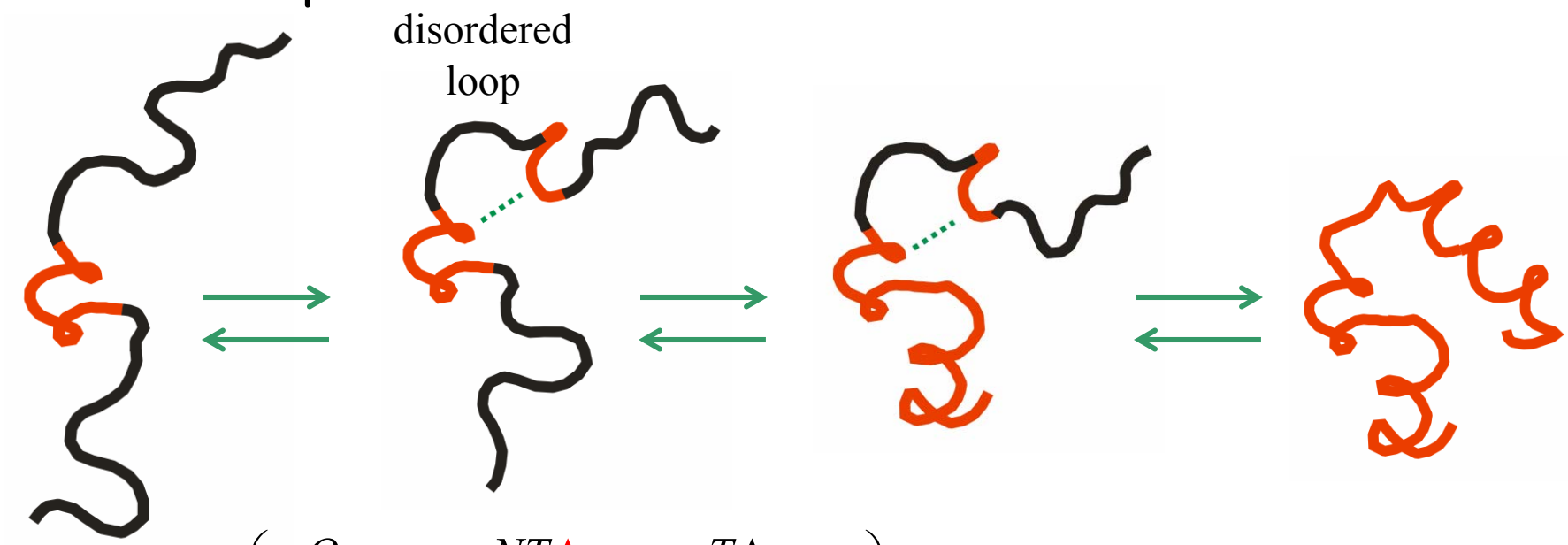


Rules of Ising-like model of Muñoz/Henry/WAE for proteins

Residues in either native or non-native conformation
Contacts only if all intervening residues in native conformation
Native structure grows in no more than two regions (DSA)
(reduces conformations from $2^{35} \sim 10^{10}$ to $\sim 10^{5}$)
Native segments can be connected by disordered loop

$$\begin{array}{l} c\,ccccnnnnnnccc\ldots \\ \quad\;\; \| \\ c\,ccccnnnccccccc\ldots \end{array}$$

Non-native interactions not explicitly considered – reflected in landscape ***D***

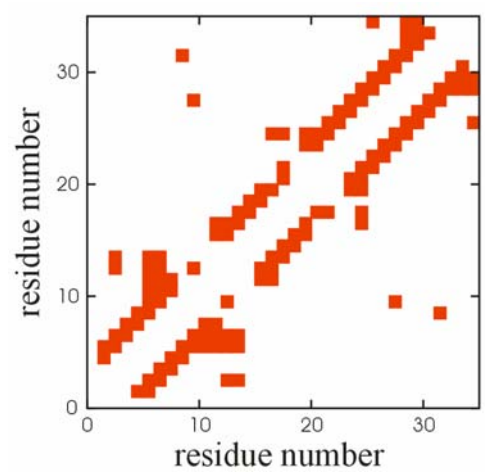


$$P_{eq} = \exp\left(-\frac{Q\varepsilon_{\text{contact}} - NT\Delta s_{\text{conf}} - T\Delta g_{\text{loop}}}{k_B T}\right)$$

"Heteropolymer": 0 or ε

Input:

I. residue-residue contact map of native structure



8 Å contacts

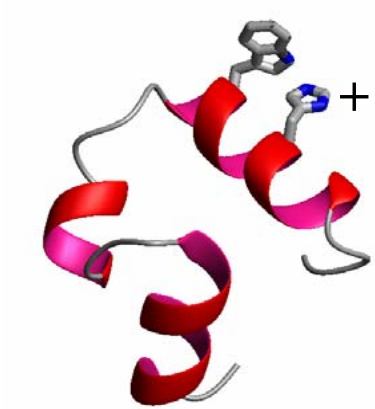
II. Three thermodynamic parameters: $\varepsilon_{\text{contact}}$, Δs_{conf}, Δg_{loop}

partition function: $Z = \sum_{i=1}^{97,769} \exp\left(-\frac{G_i}{k_B T}\right) = \sum_{i=1}^{97,769} \exp\left(-\frac{Q_i \varepsilon_{\text{contact}} - N_i T \Delta s_{\text{conf}} - T \Delta g_{i,\text{loop}}}{k_B T}\right)$

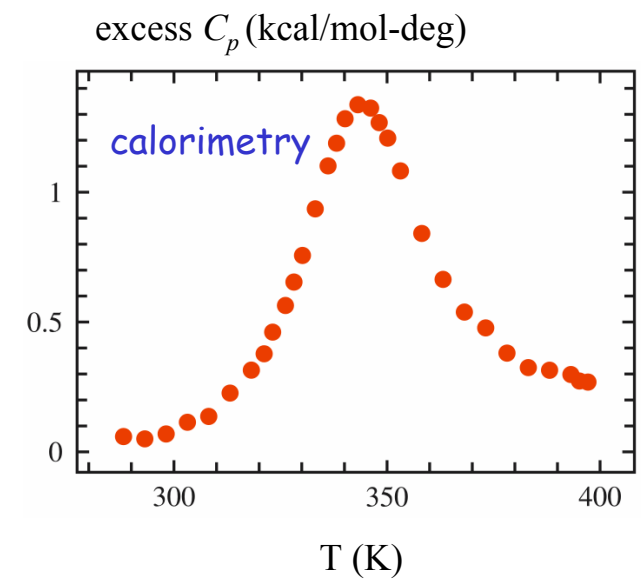
$$\Delta g_{\text{loop}} = -k_B T \ln\left(\frac{\int_a^{a+\delta a} P(r)dr}{\int_a^{l_c} P(r)dr}\right), \quad P(r) = \frac{4\pi N r^2}{l_c^2\left(1-(r/l_c)^2\right)^{9/2}} \exp\left(\frac{3l_c}{4\boxed{l_p}\left(1-(r/l_c)^2\right)}\right)$$

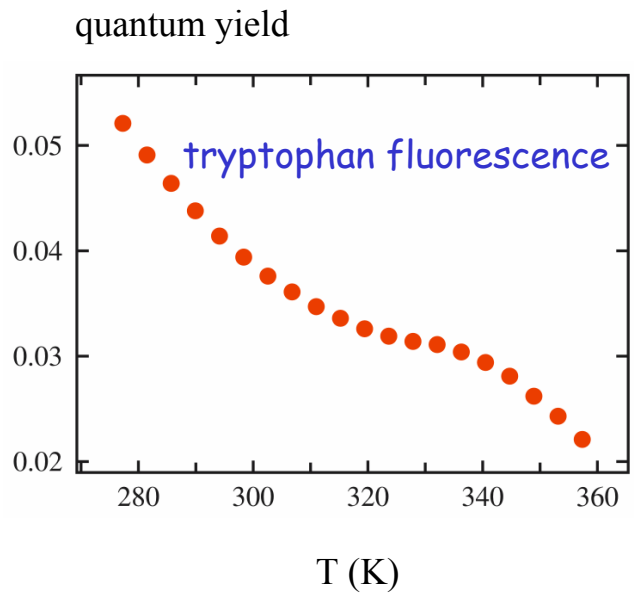
D. Thirumalai

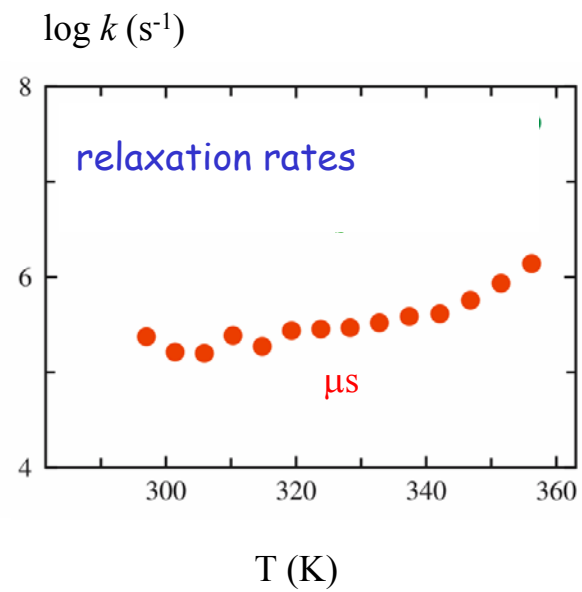
III. two fluorescence parameters to describe temperature dependence of fractional tryptophan quantum yield at contact with protonated histidine: $f_0 + \beta(T-T_0)^2$

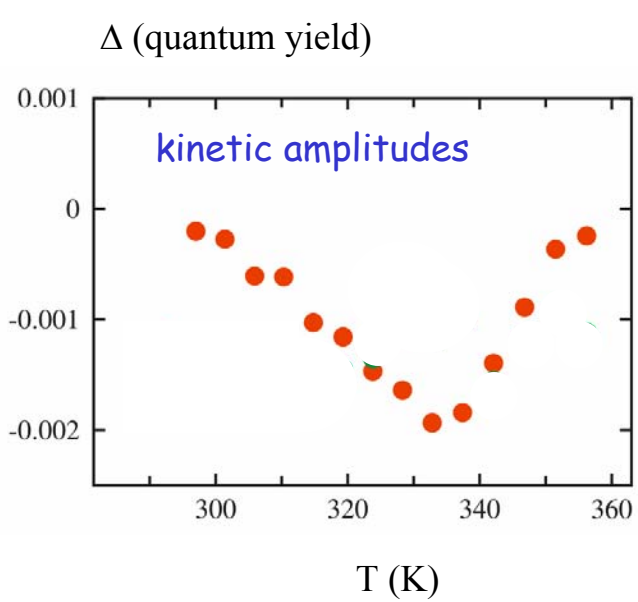


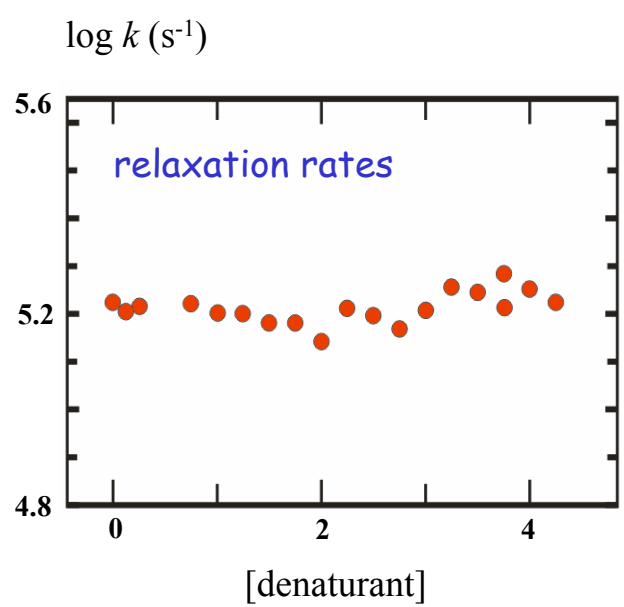
Data to be calculated by theoretical model

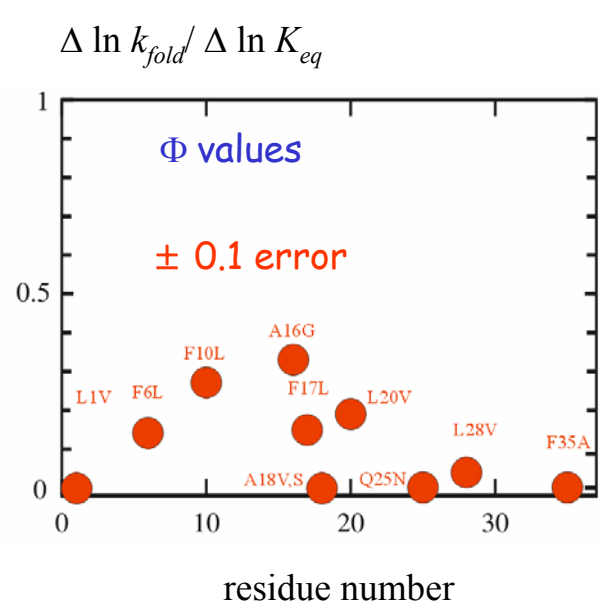








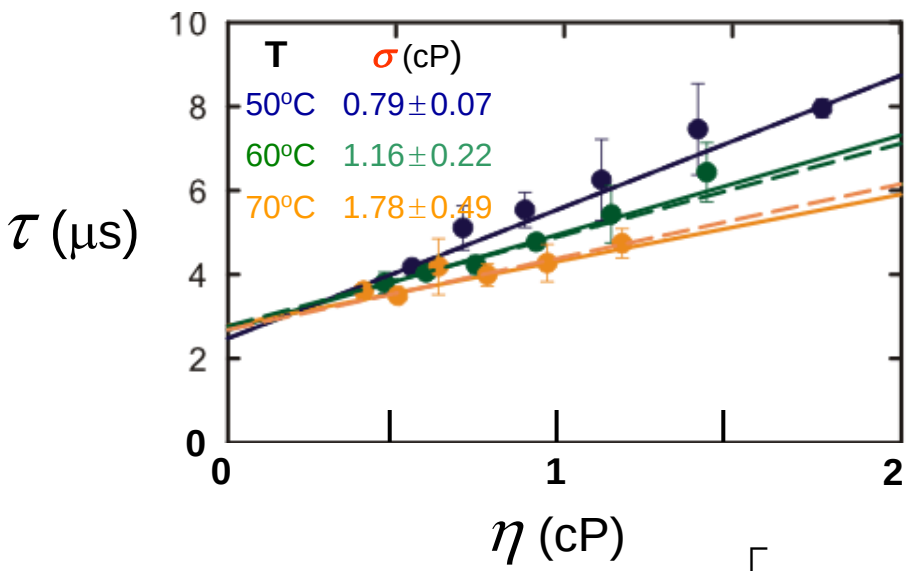


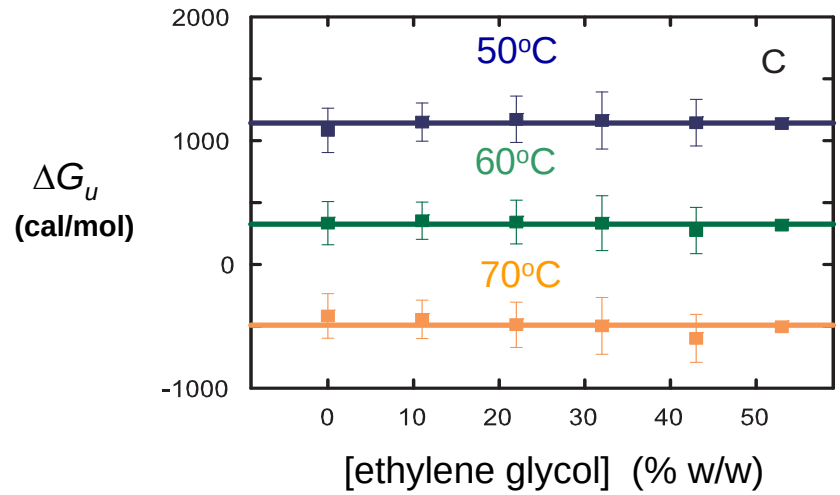


Data to be calculated by theoretical model (cont.)

Cellmer, Henry, Hofrichter, WAE, PNAS 2008

relaxation rates





$$\frac{1}{\tau} = \frac{1}{\tau_f} + \frac{1}{\tau_u} = \frac{D^{\ddagger}\omega^{\ddagger}}{2\pi k_B T}\left[\omega_u \exp\left(-\frac{\Delta G_f^{\ddagger}}{k_B T}\right) + \omega_f \exp\left(-\frac{\Delta G_u^{\ddagger}}{RT}\right)\right]$$

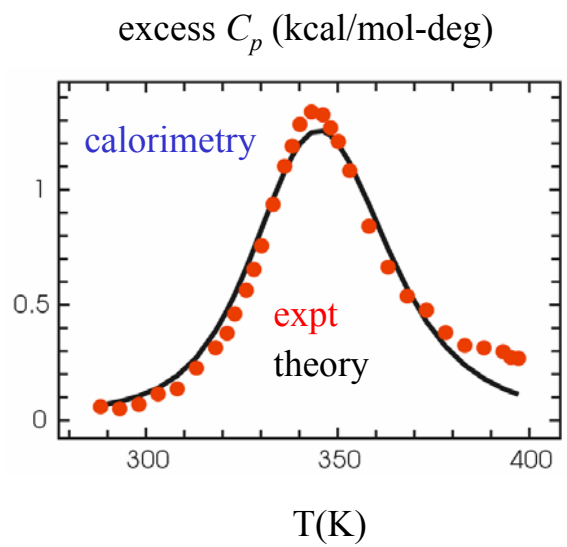
$$\tau \sim \frac{\sigma + \eta}{B}$$

continuous curves

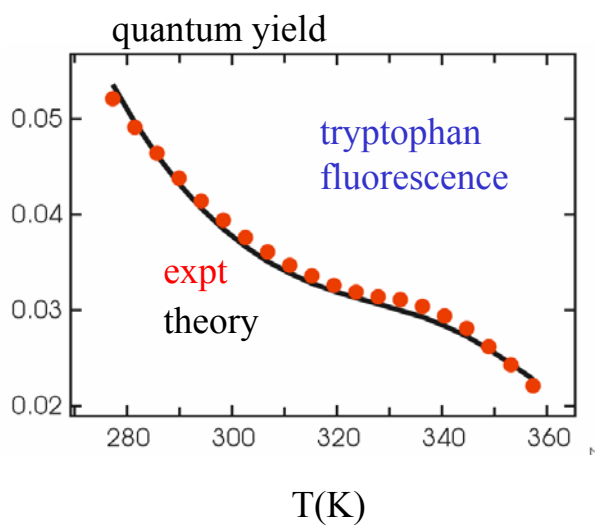
Internal viscosity, σ, increases with increasing T

Ansari, Jones, Henry, Hofrichter, WAE, Science 1992

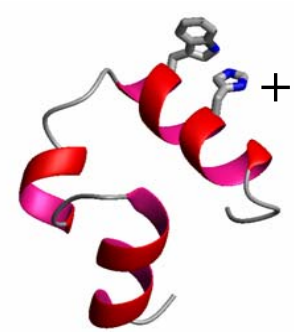
Calculation of equilibrium properties (Δs_{conf} = -3.8 eu)



$$C_p^{ex} = \frac{dH}{dT} \approx \frac{dE}{dT} = \frac{d}{dT}\left(k_{\mathrm{B}}T^2 \frac{d\ln Z}{dT}\right)$$

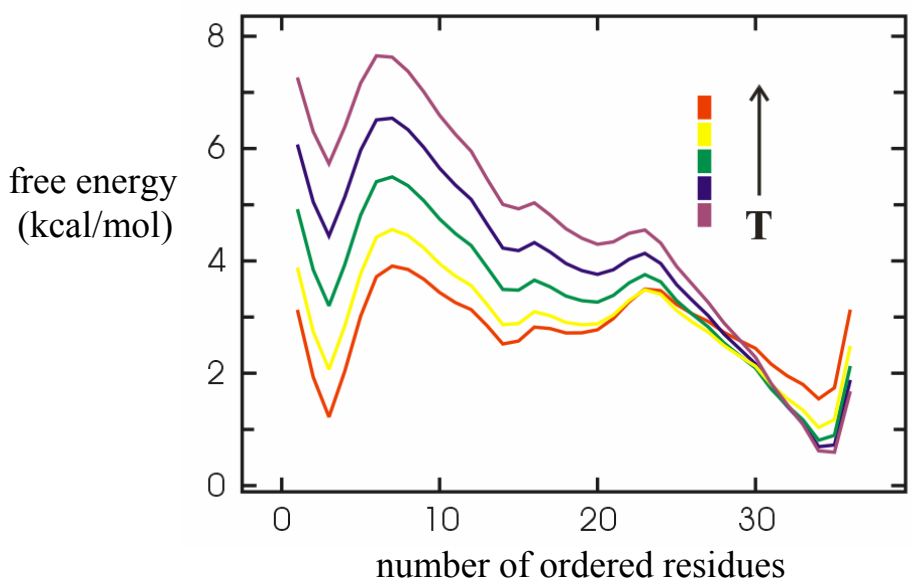


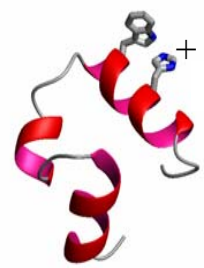
$$\Phi = \frac{\sum_{i=1}^{97,769} f_i\, \phi_{\mathrm{ref}} \exp\left(-\frac{G_i}{k_{\mathrm{B}}T}\right)}{Z}$$

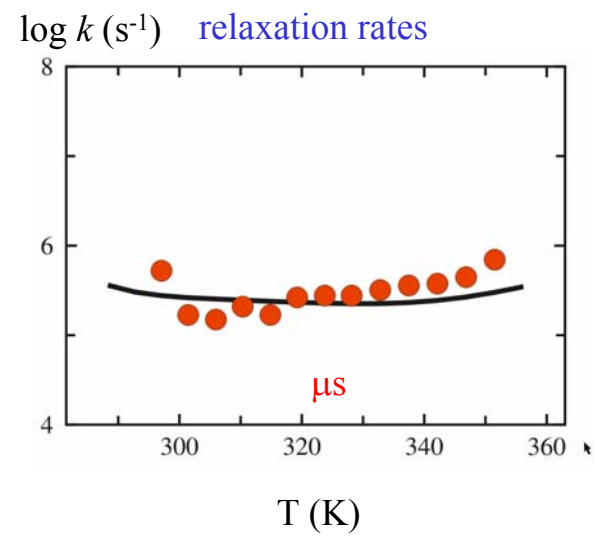


$$f_i = \begin{cases} f_0 + \beta(T - T_0)^2 & \text{if W-H contact formed} \\ 1 & \text{otherwise} \end{cases}$$

Calculation of kinetic properties: diffusion on 1D free energy surface



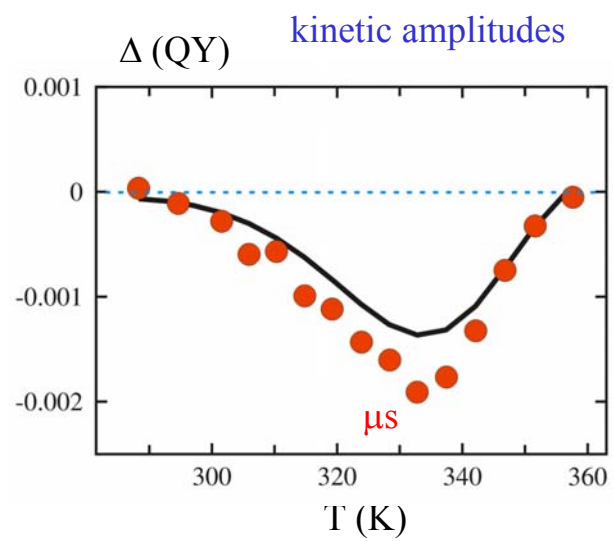


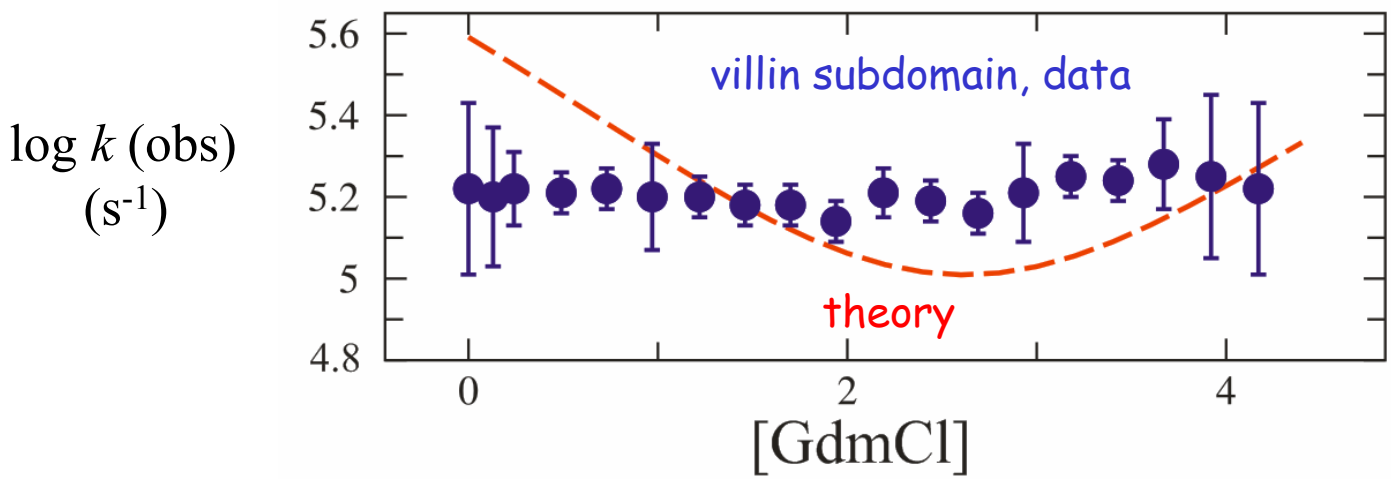


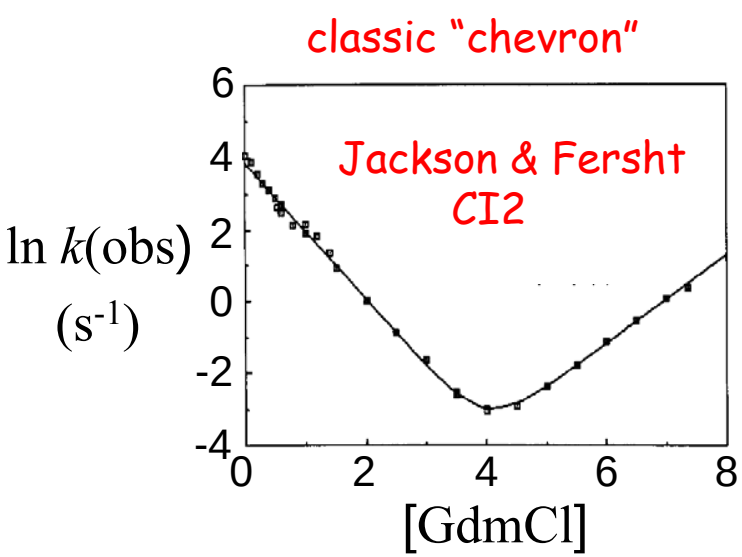
Simulate diffusion by hopping along reaction coordinate with kinetic rule: γ and its activation energy are only additional parameters

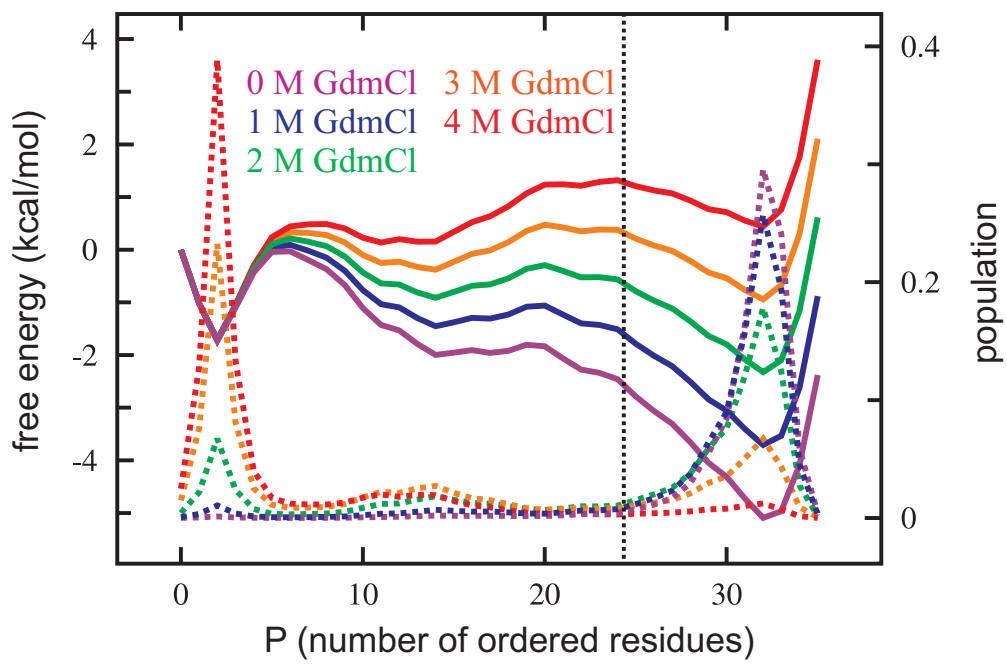
$$k_{ij} = \gamma K_{ij}^{1/2} = \gamma \exp\left(-\frac{\Delta G_{ij}}{2k_B T}\right)$$

$$\frac{d}{dt}\begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_{N-1} \\ p_N \end{pmatrix} = \begin{pmatrix} -k_{0,1} & k_{1,0} & 0 & \cdots & \cdots & 0 \\ k_{0,1} & -(k_{1,0}+k_{1,2}) & k_{2,1} & \cdots & \cdots & 0 \\ 0 & k_{1,2} & -(k_{2,1}+k_{2,3}) & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & k_{N-1,N-2} & 0 \\ \vdots & \vdots & \vdots & \vdots & -(k_{N-1,N}+k_{N-1,N-2}) & k_{N,N-1} \\ 0 & \cdots & \cdots & 0 & k_{N-1,N} & -k_{N,N-1} \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_{N-1} \\ p_N \end{pmatrix}$$





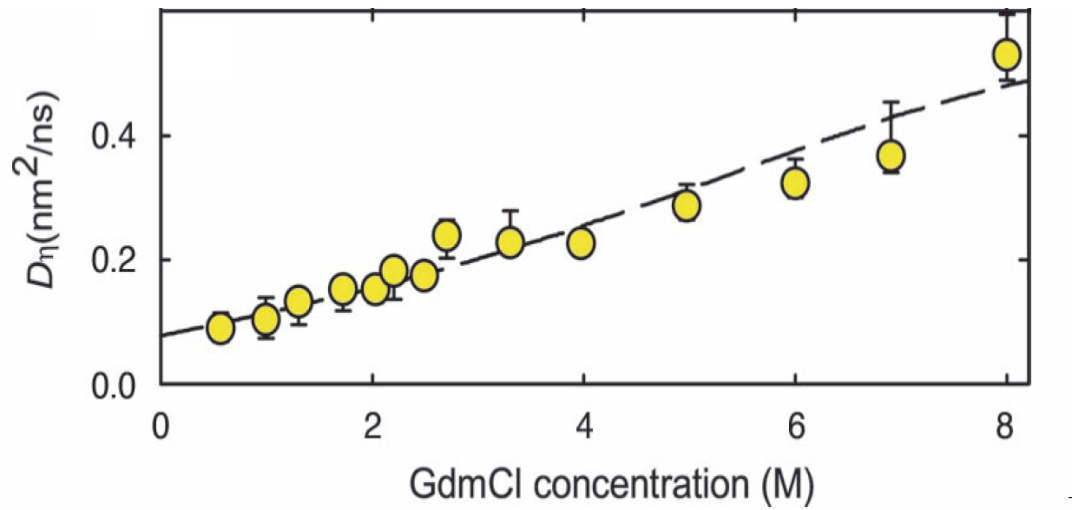




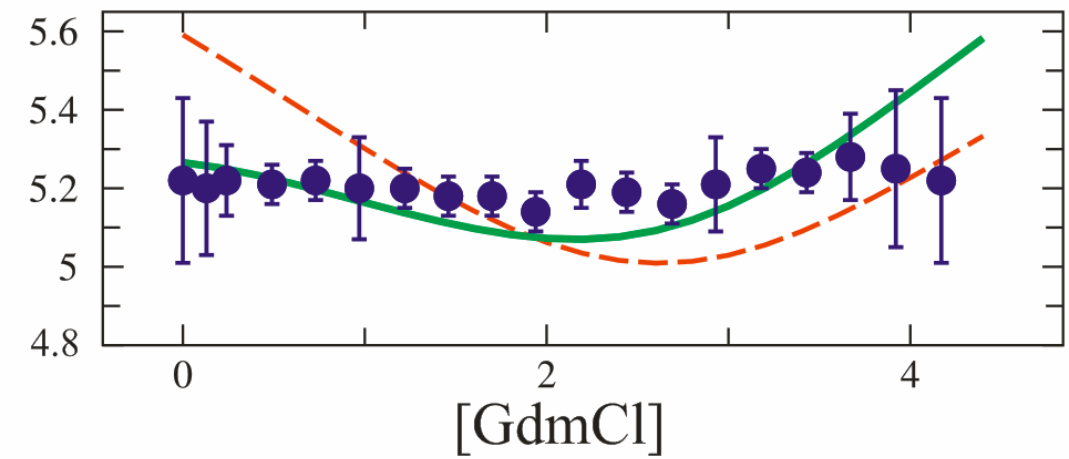
No denaturant dependence because of insensitivity of barriers to energy change and because major barrier shifts

Folding barrier increases with denaturant, but so does diffusion coefficient!!

B. Schuler (U. Zürich) *PNAS* 2007

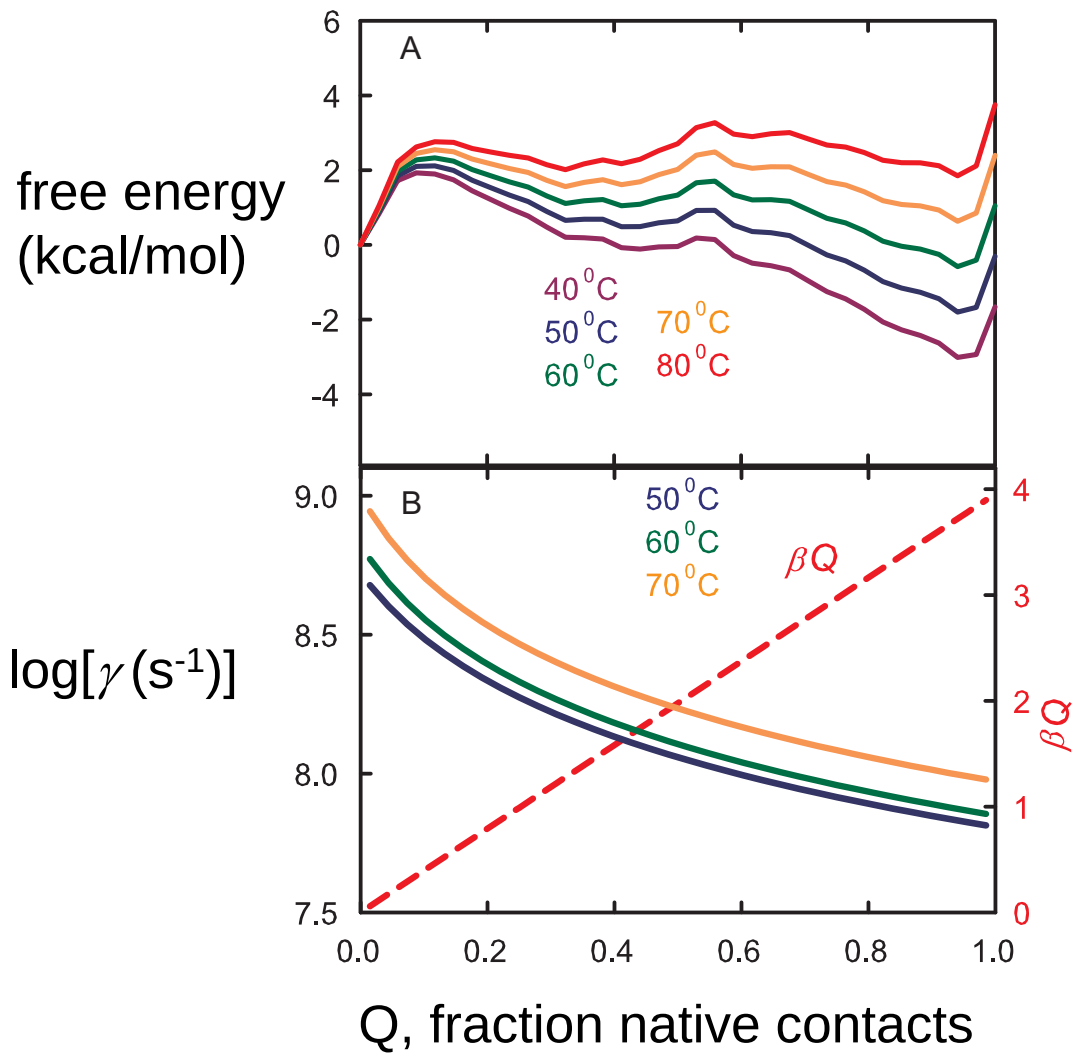


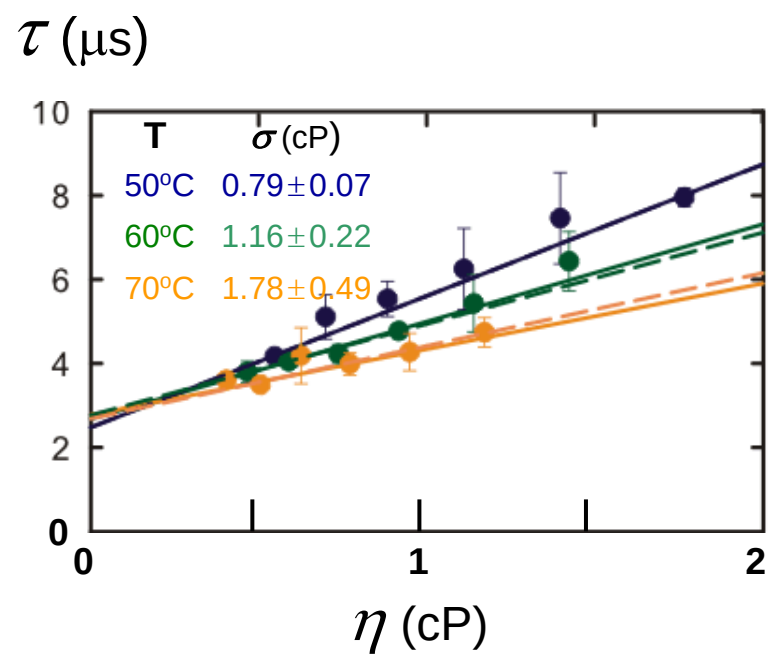
log k (obs) s^{-1} with Schuler GdmCl dependence



Introduce position dependent diffusion coefficient to explain temperature dependence of internal viscosity

$$\gamma(Q,T) = \frac{\Gamma(T)}{\alpha + \beta Q + \eta(T)}$$





points data

$$\tau = \frac{\sigma + \eta}{B}$$ continuous

model dashed

Do the reaction coordinates

P (number of native residues)
and
Q (fraction of native contacts)

reproduce transition state results based on responses to local perturbation, i.e. points mutations

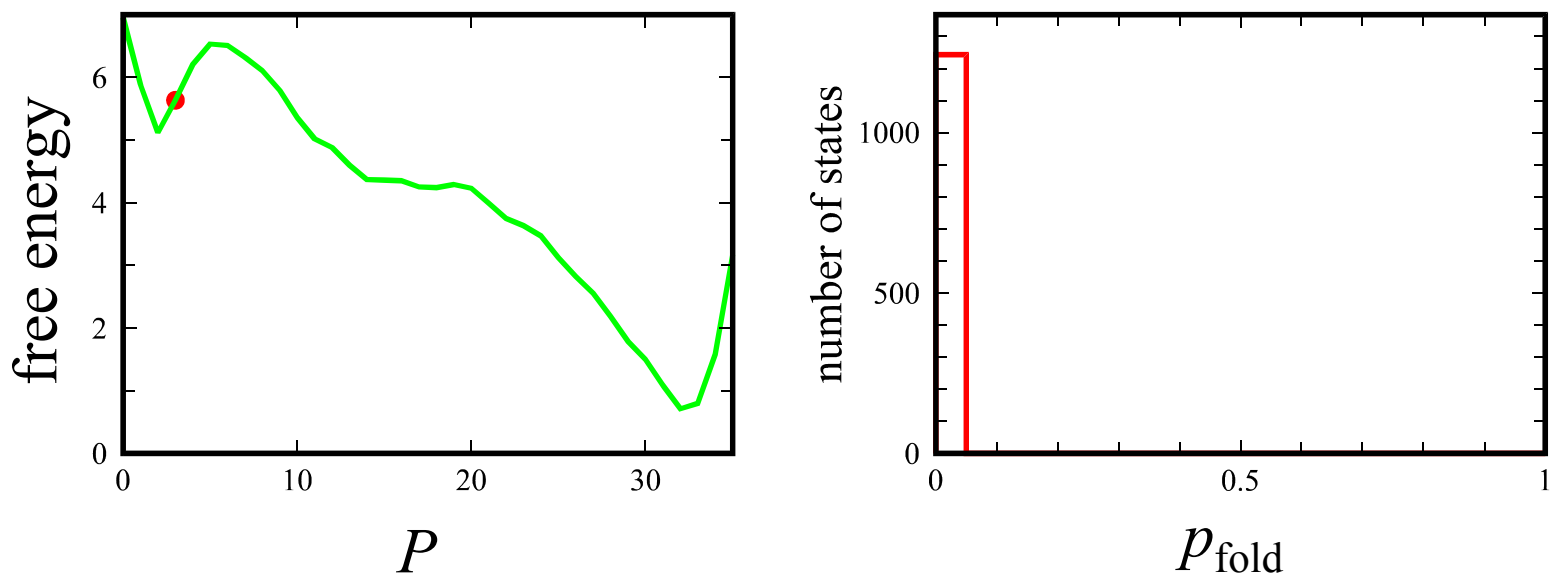
Du, Pande, Grosberg, Tanaka, & Shakhnovich, *J. Chem. Phys.* 1998

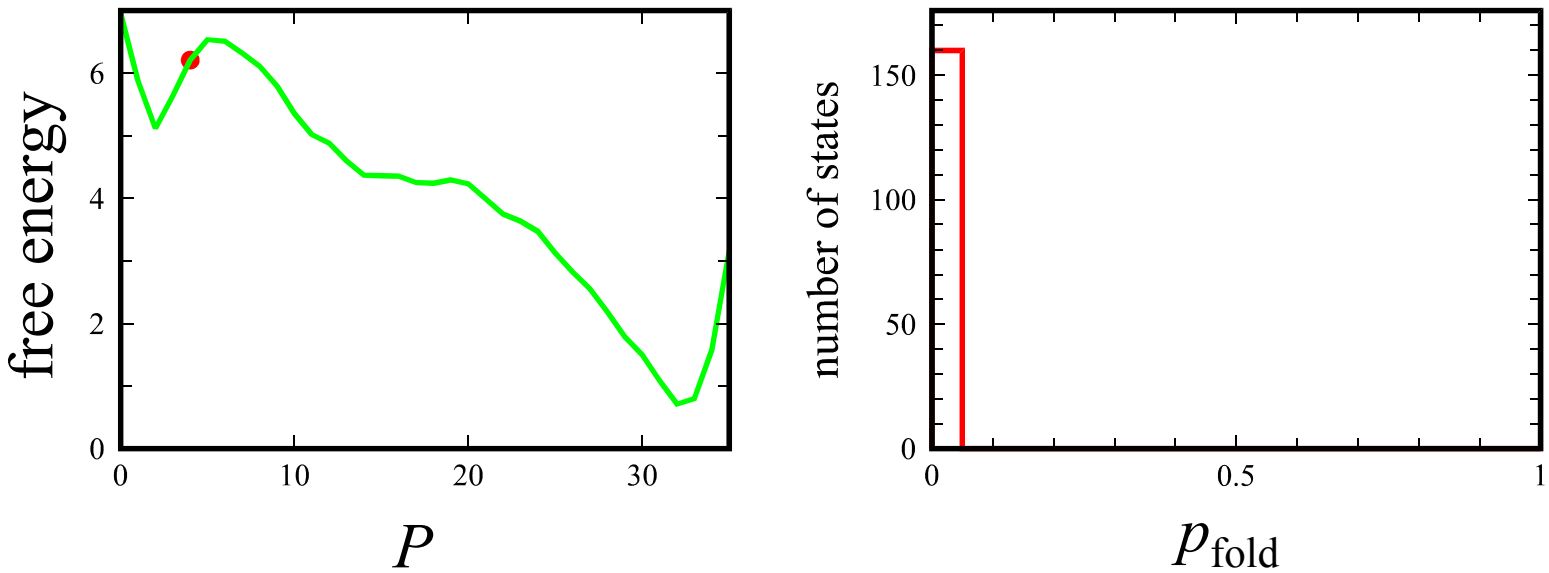
p_{fold} is rigorous criterion for determining whether a structure belongs to the transition state ensemble

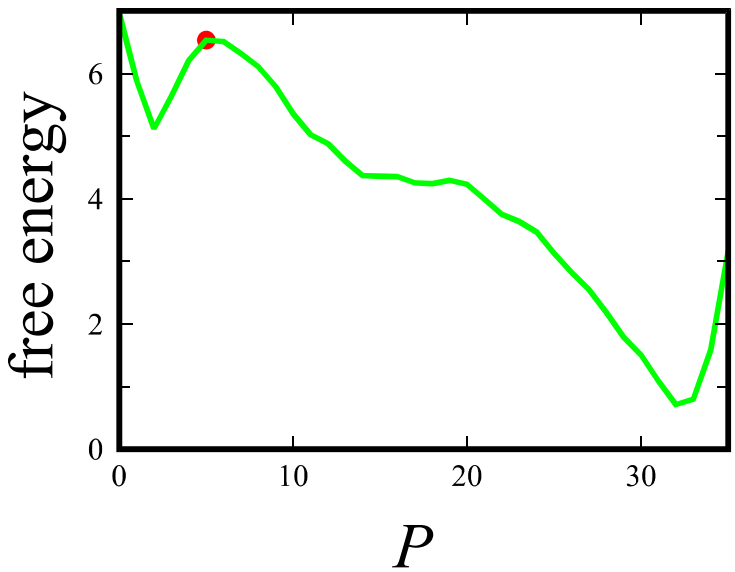
p_{fold} is defined as the probability of reaching the folded state for the first time before reaching the unfolded state for the first time. If p_{fold} is ~ $\frac{1}{2}$, structure belongs to transition state ensemble.

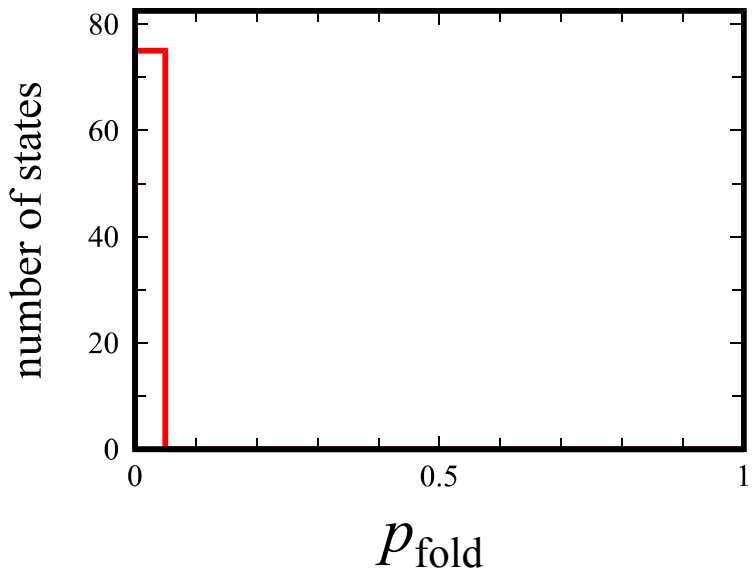
p_{fold} for each of the 97,769 species of model can be directly calculated from the rate matrix

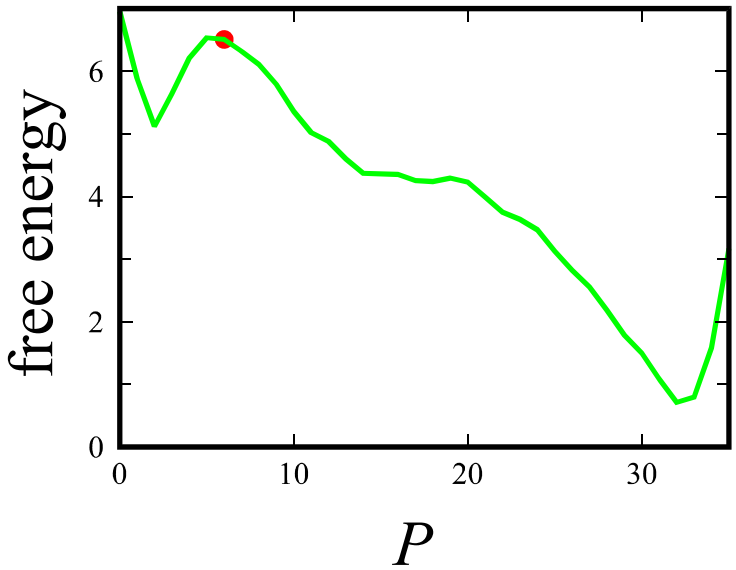
$$\mathbf{K}^{\mathbf{T}}\boldsymbol{\sigma} = \mathbf{0}$$

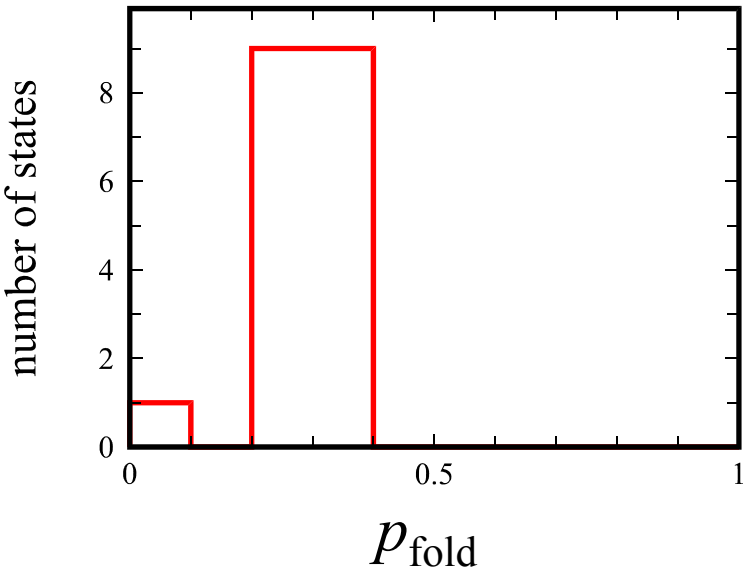

free energy
0
2
4
6
0
10
20
30
P
number of states
0
500
1000
0
0.5
1
p_{fold}

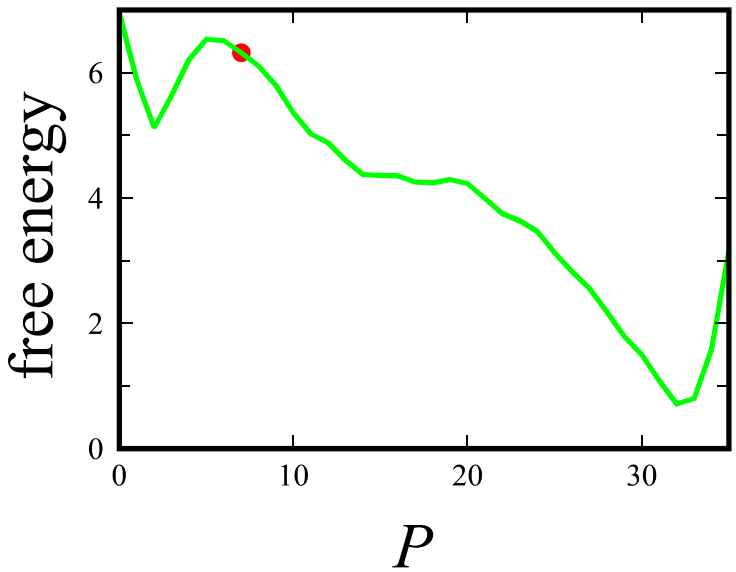

free energy
0
2
4
6
0
10
20
30
P
number of states
0
50
100
150
0
0.5
1
p_{fold}

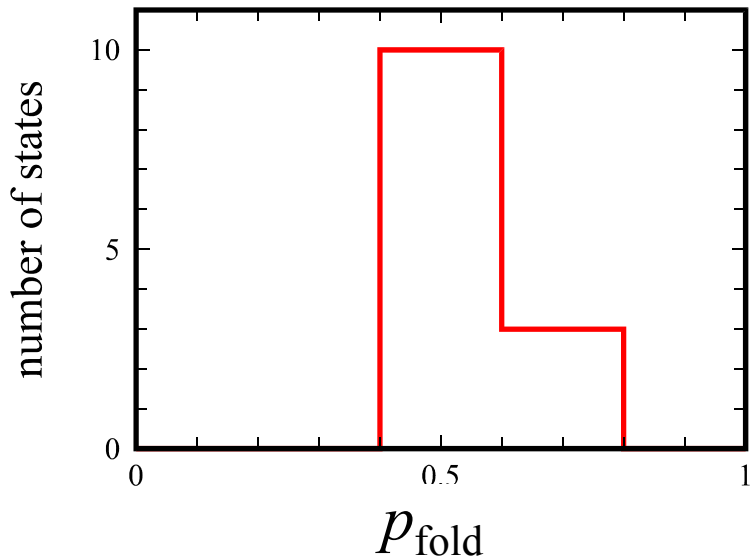

free energy
0
2
4
6
0
10
20
30
P

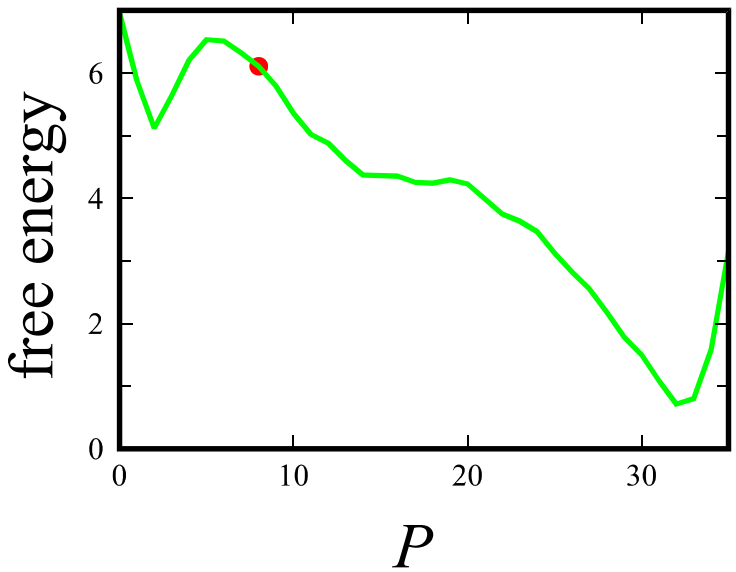

number of states
0
20
40
60
80
0
0.5
1
p_{fold}

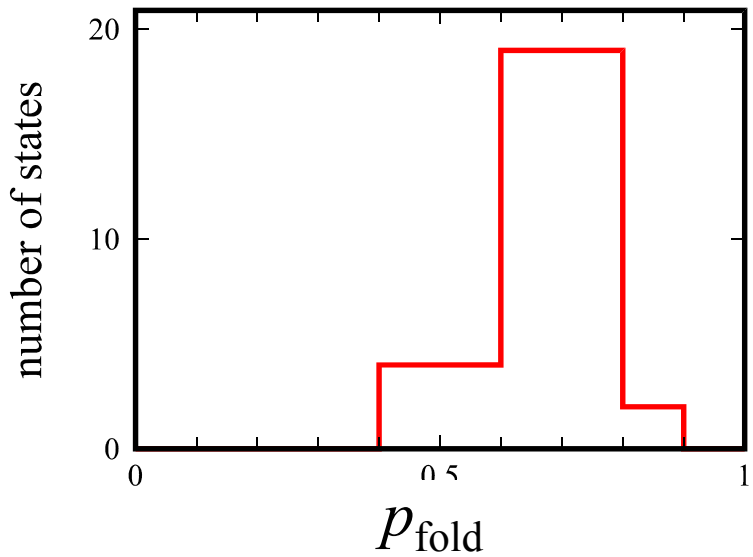

free energy
0
2
4
6
0
10
20
30
P

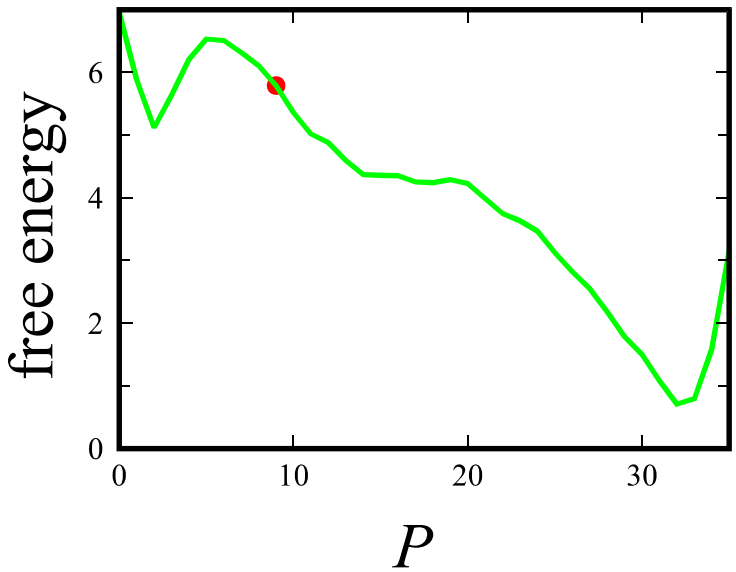

number of states
0
2
4
6
8
0
0.5
1
p_{fold}

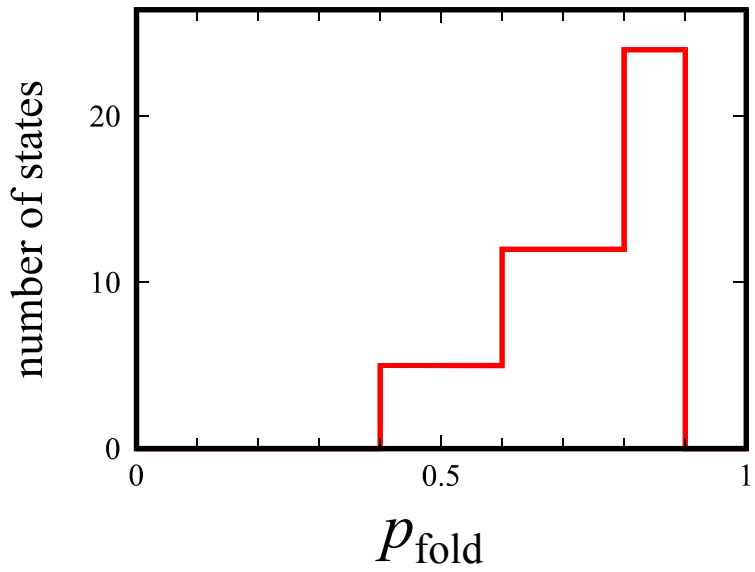

free energy
0
2
4
6
0
10
20
30
P

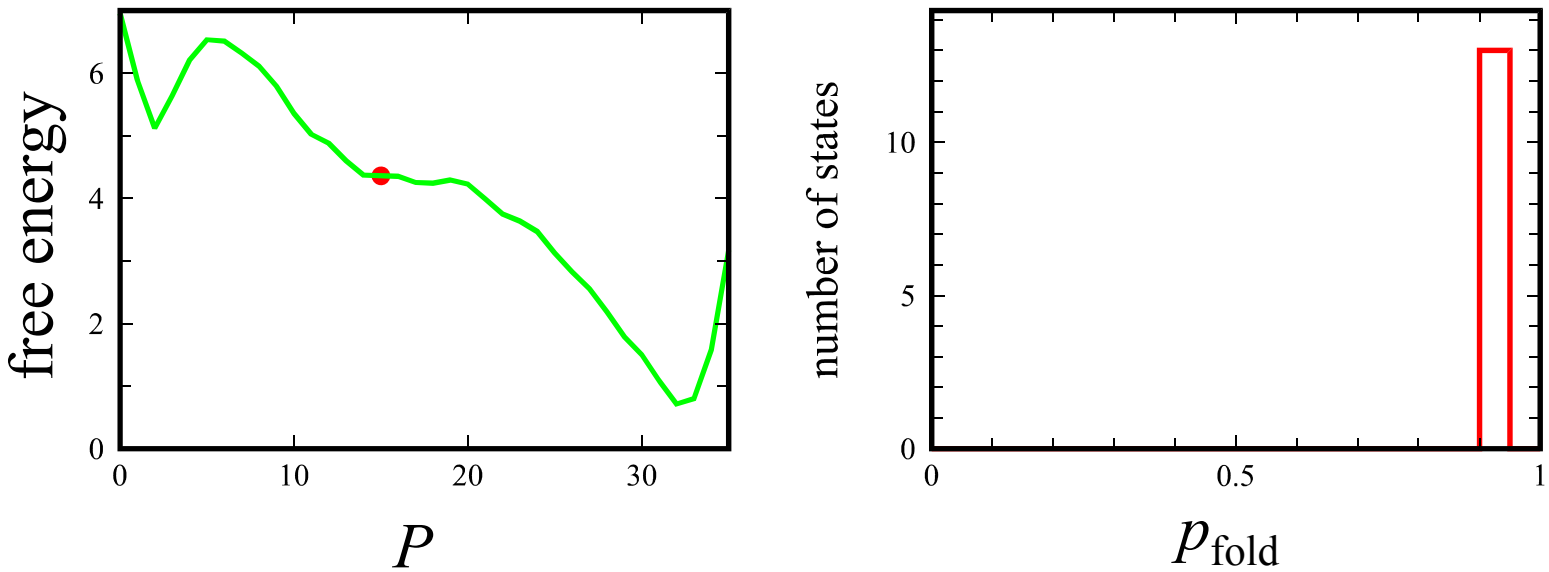

number of states
0
5
10
0
0.5
1
p_{fold}

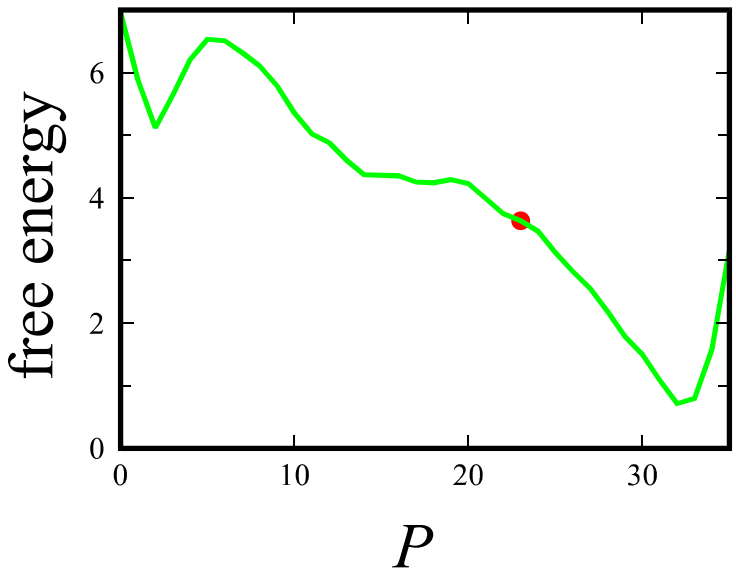

free energy
0
2
4
6
0
10
20
30
P

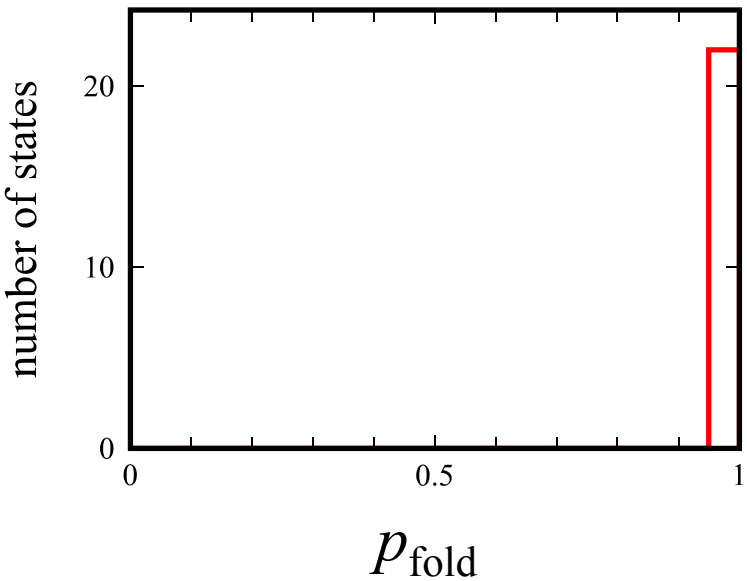

number of states
0
10
20
0
0.5
1
p_{fold}


free energy
0
2
4
6
0
10
20
30
P

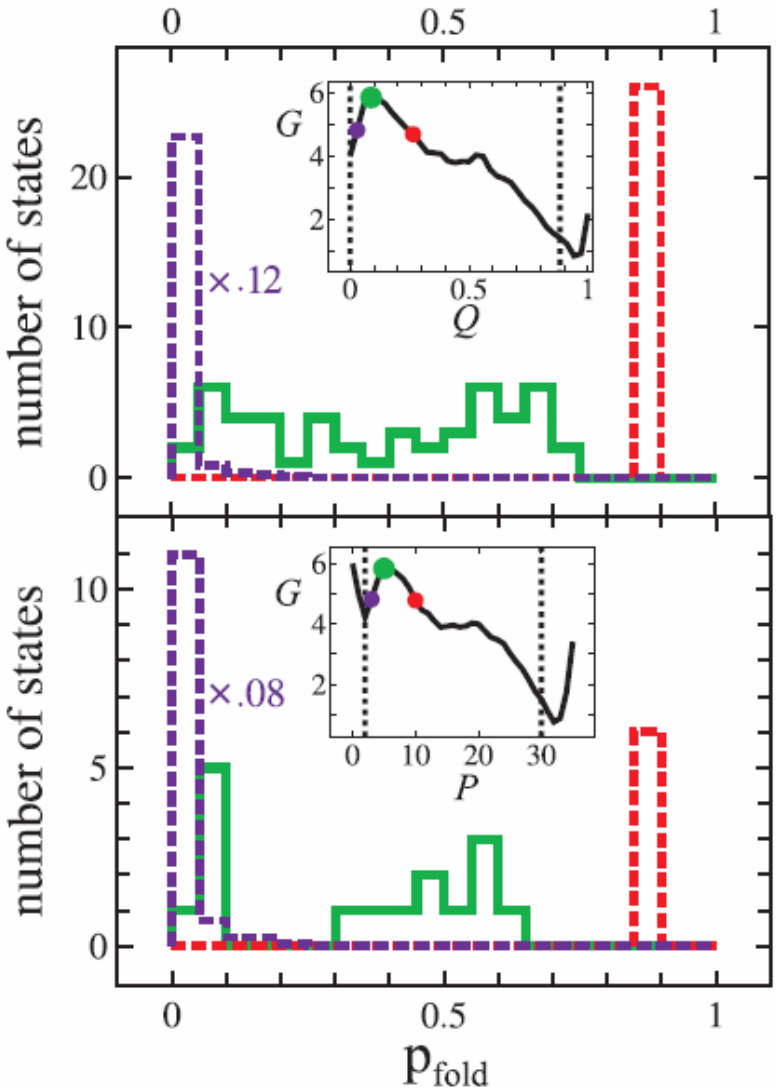

number of states
0
10
20
0
0.5
1
p_{fold}


free energy
0
2
4
6
0
10
20
30
P
number of states
0
5
10
0
0.5
1
p_{fold}


free energy
0
2
4
6
0
10
20
30
P


number of states
0
10
20
0
0.5
1
p_{fold}

Splitting probability distributions for all states <2k_BT from most stable state at each value of ***P*** and ***Q***.



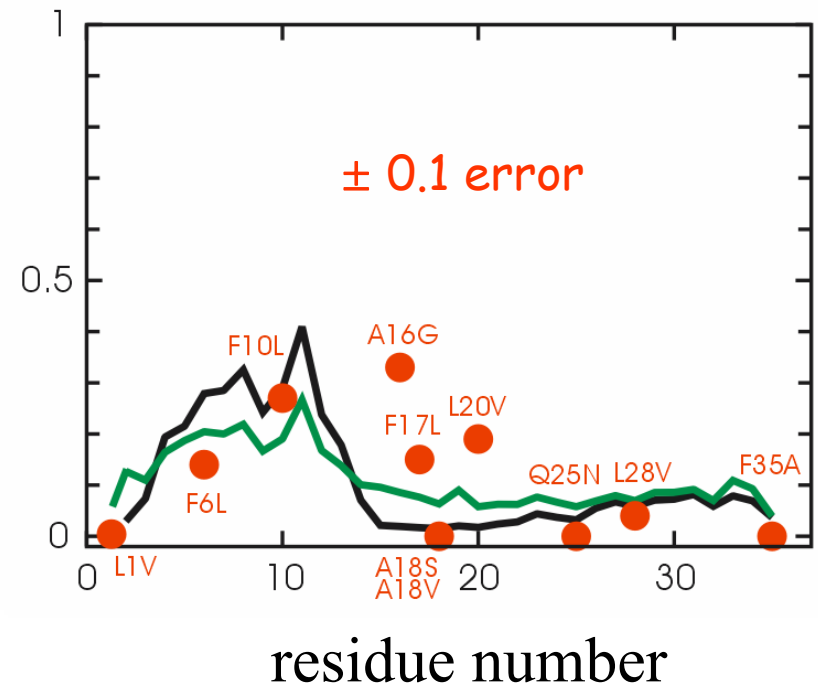
Conformations at top of free energy barrier belong to TS, therefore ***P*** and ***Q*** appear to be good reaction coordinates

Calculation of Φ values (response to a local perturbation)

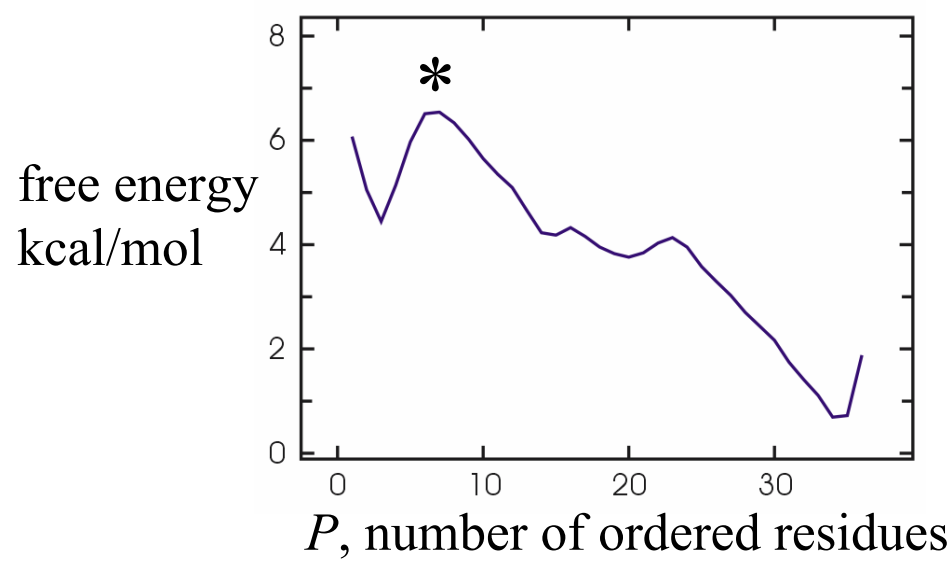
Interpretation: fraction of native contacts for conformations of the transition state (i.e., top of free energy barrier)

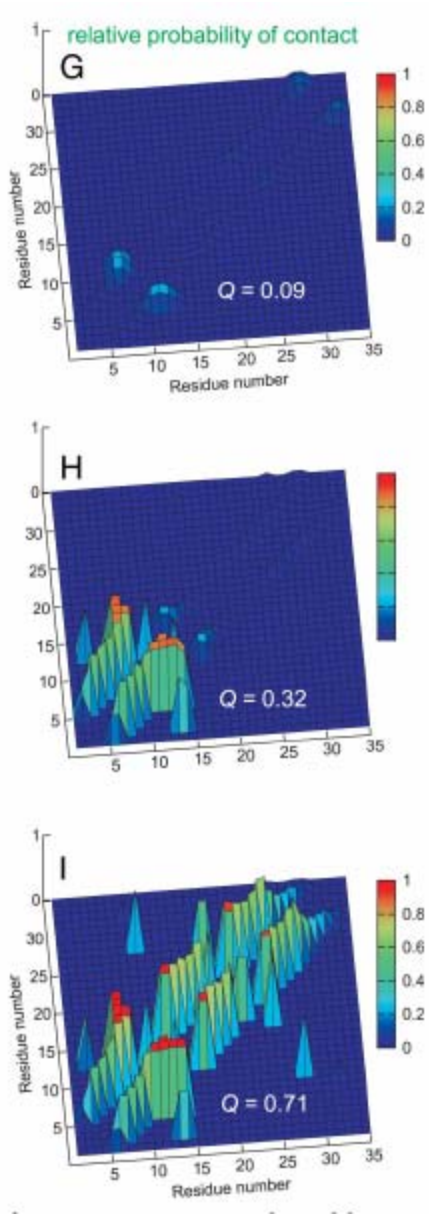
$$\Phi_j = \frac{\sum_{i*} \frac{Q_{j,i*}}{Q_{j,native}} \exp\left(-\frac{G_{i*}}{k_B T}\right)}{\sum_{i*} \exp\left(-\frac{G_{i*}}{k_B T}\right)}$$

$$\Phi \equiv \frac{\Delta \ln k_{fold}}{\Delta \ln K_{eq}}$$



More rigorous: calculate **experimental result**, i.e. $\Delta \ln k_{fold} / \Delta \ln K_{eq}$ change contact energy of mutated; changes free energy surface; calculate new equilibrium constant; calculate new rate coefficient





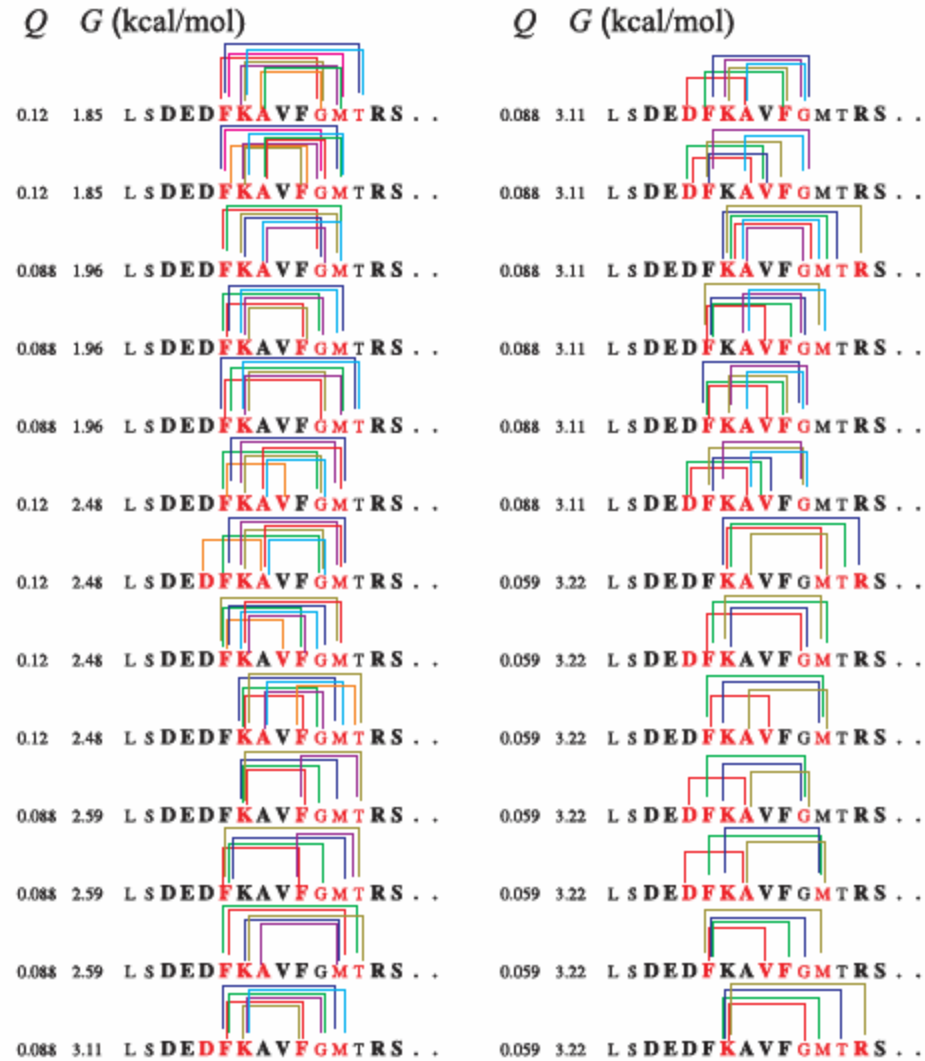


Fig. S8. Most populated microstates of transition-state ensemble for Q reaction coordinate at 310 K. The first 15 residues of the sequence are shown. Red letters correspond to residues in the native conformation; black letters correspond to residues in the nonnative conformation; large letters in bold type are residues that are helical in the native structure according to the criterion of Kabsch and Sander (12). Interresidue contacts in each microstate are indicated by continuous lines. For the Q reaction coordinate, the binning was twofold, so the $Q = 0.059$ bin contains microstates with either three or four contacts, the $Q = 0.088$ bin contains microstates with either five or six contacts, and the $Q = 0.12$ bin contains residues with either seven or eight contacts. The transition state ensemble of microstates was defined to include those microstates with (*i*) a p_{fold} between 0.4 and 0.6, (*ii*) with values of the reaction coordinate (i.e., $Q = 0.059$ and $Q = 0.12$) for which the free energy is within 1 k_BT of the free energy at the maximum ($Q = 0.088$), and (*iii*) with free energies <2 k_BT above the minimum free energy for all microstates at the specified value of Q. For $Q = 0.059$, the seven microstates shown contain 2% of the equilibrium population of all microstates, including those with $0.6 < p_{fold} < 0.4$; the corresponding numbers for $Q = 0.088$ are 13 microstates and 12% of the equilibrium population, and for $Q = 0.12$, six microstates and 16% of the equilibrium population.

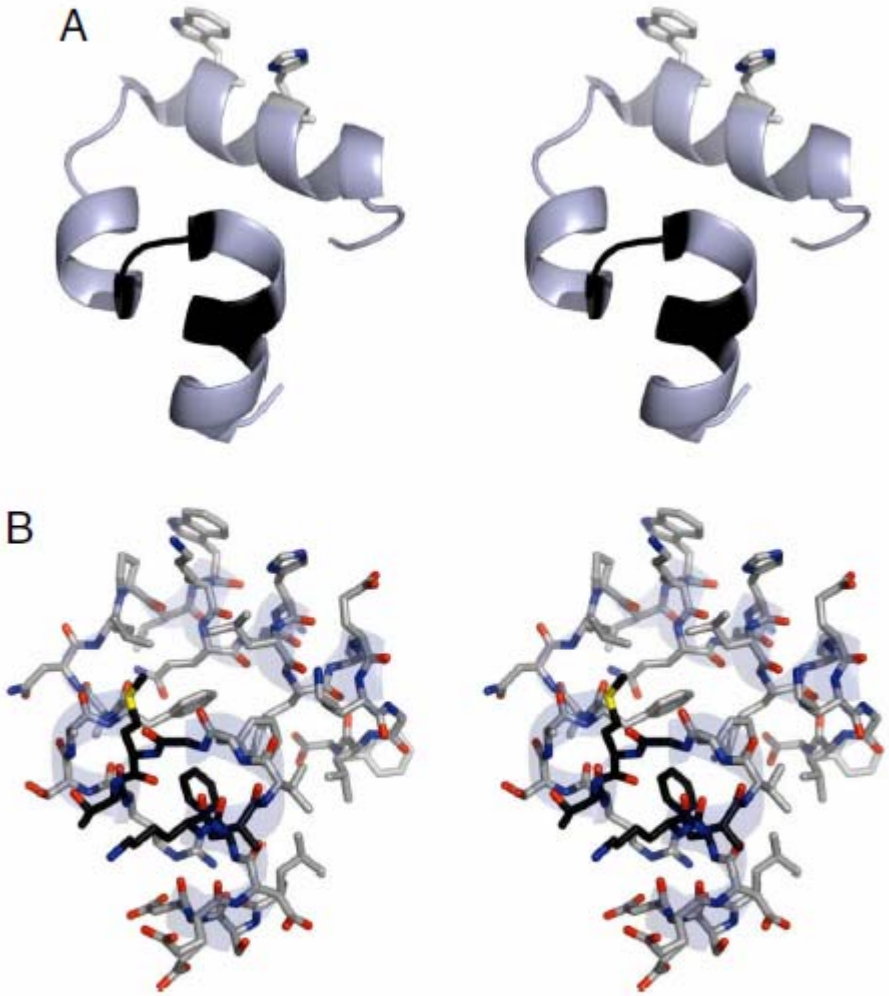
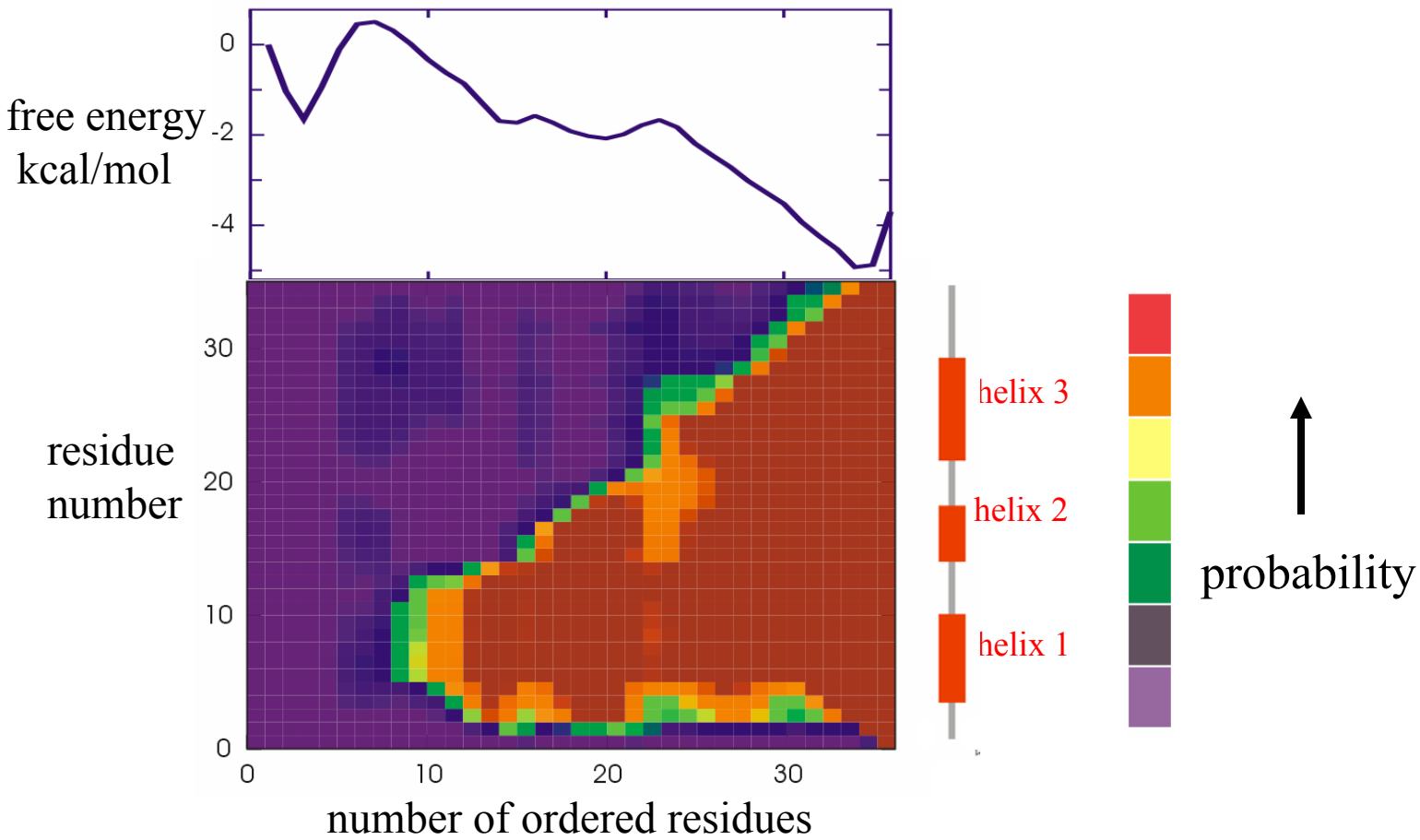


Fig. 1. Structure of villin subdomain solved by x-ray diffraction (PDB 1WY4) (2). Ribbon diagram of backbone showing the side chains of W23 and H27 (*A*) and structure with all nonhydrogen atoms (*B*). Residues F6,K7,A8,G11,M12,T13 are shown in black because their contacts contribute most to the stability of the most populated microstates of the transition state ensemble at 310 K (see Figs. S8 and S9).

Mechanistic insight - order of secondary structure formation

probability that residue is in native conformation for a given value of the reaction coordinate



Summary

A very simple, **coarse-grained** statistical mechanical model based on the contact map of the **native** structure that only considers the tradeoff between conformational entropy and stabilizing contacts (with same energy and entropy for every residue) is remarkably successful in quantitatively explaining a wide variety of experimental data, i.e. equilibrium calorimetry, fluorescence, circular dichroism; relaxation rates as a function of temperature, denaturant, and viscosity, kinetic amplitudes; and effects of mutations on folding rates (Φ values).

Next Steps

Test contiguous sequence approximation with Langevin simulations of Gō bead model – currently being carried out by Robert Best

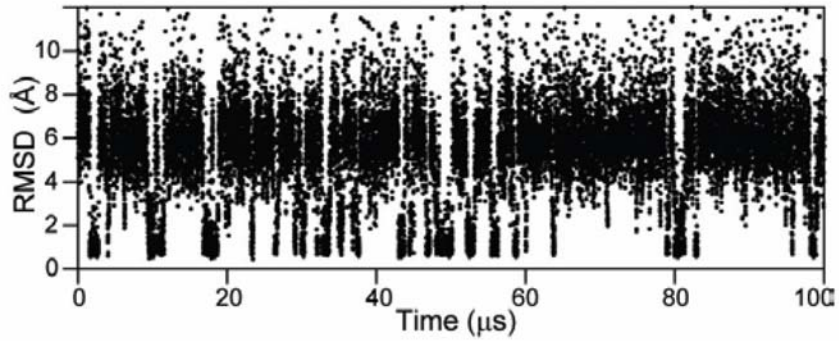
Compare predicted mechanism with (David E. Shaw) MD trajectories in explicit water using his super-computer "Anton", hard wired for MD

Use single molecule FRET measurements to observe distribution of transition paths – current major focus of my lab

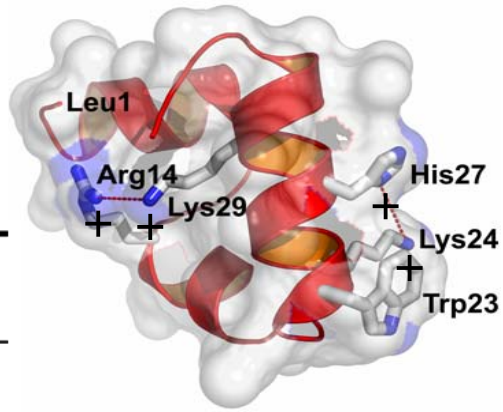
S. Piana, K. Lindorff-Larsen, and D.E. Shaw

(*Biophysical Journal– Biophysical Letters*, 2011)

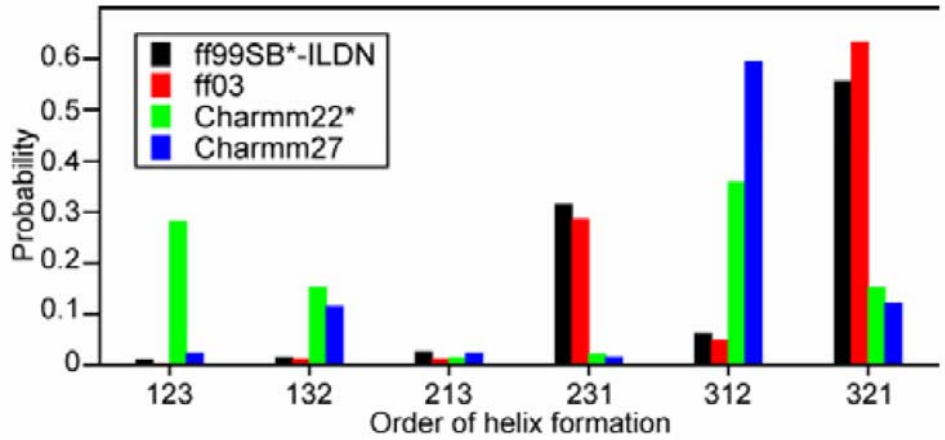
Shaw's "Anton"



Double norleucine mutant "supervillin"



	T (K)	ΔG_f (kcal mol^{-1})	ΔH_f (kcal mol^{-1})	Folding time (μs)	% of helix 1/2/3
ff03	390	0.21±0.1	9.7±1	0.8±0.1	30/52/85
ff99SB*-ILDN	380	0.70±0.1	19.7±1	3.0±0.4	22/17/59
CHARMM27	430	0.51±0.1	19.3±0.4	0.9±0.1	73/33/90
CHARMM22*	360	1.0±0.2	17.0±1	2.6±0.5	41/9/44
Experiment:	**361**	**0**	**25 (+2,-5)**	**0.7**	



What is pathway distribution predicted by model?

Very preliminary results

Stochastic kinetic simulations (10^6 trajectories from 667 $P = 2$ states)

The rate for each possible transition is determined from a linear free energy relation, i.e. $k = \gamma K^{\alpha}$ for, with the same $\gamma = 2 \times 10^8\ s^{-1}$ and same exponent $\alpha = 1/2$ for each transition.

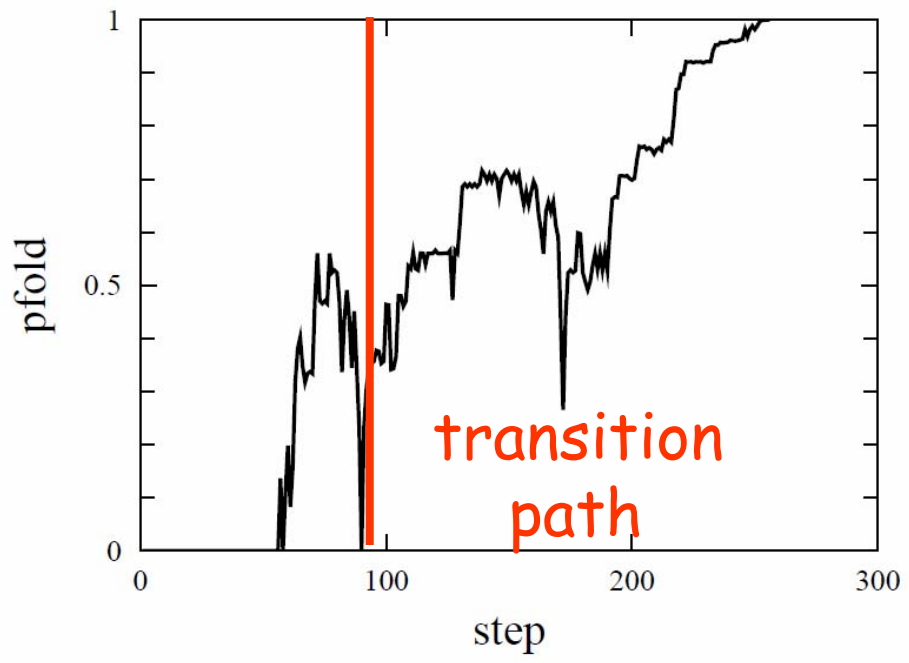
The length of a step is given by: $\tau_i = \left(\sum_j k_{ij} \right)^{-1}$

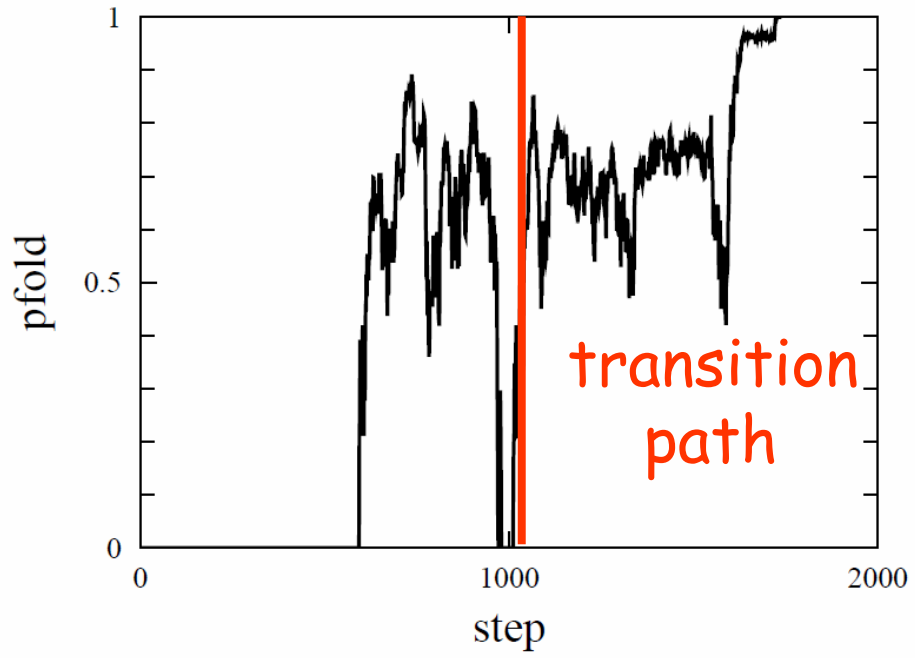
The probability of a step (...cnnc...→ ..cnnn..) is its relative rate

The rate of loop formation is taken from the experimentally-determined empirical expression for the rate of loop formation (tryptophan triplet quenching experiments)

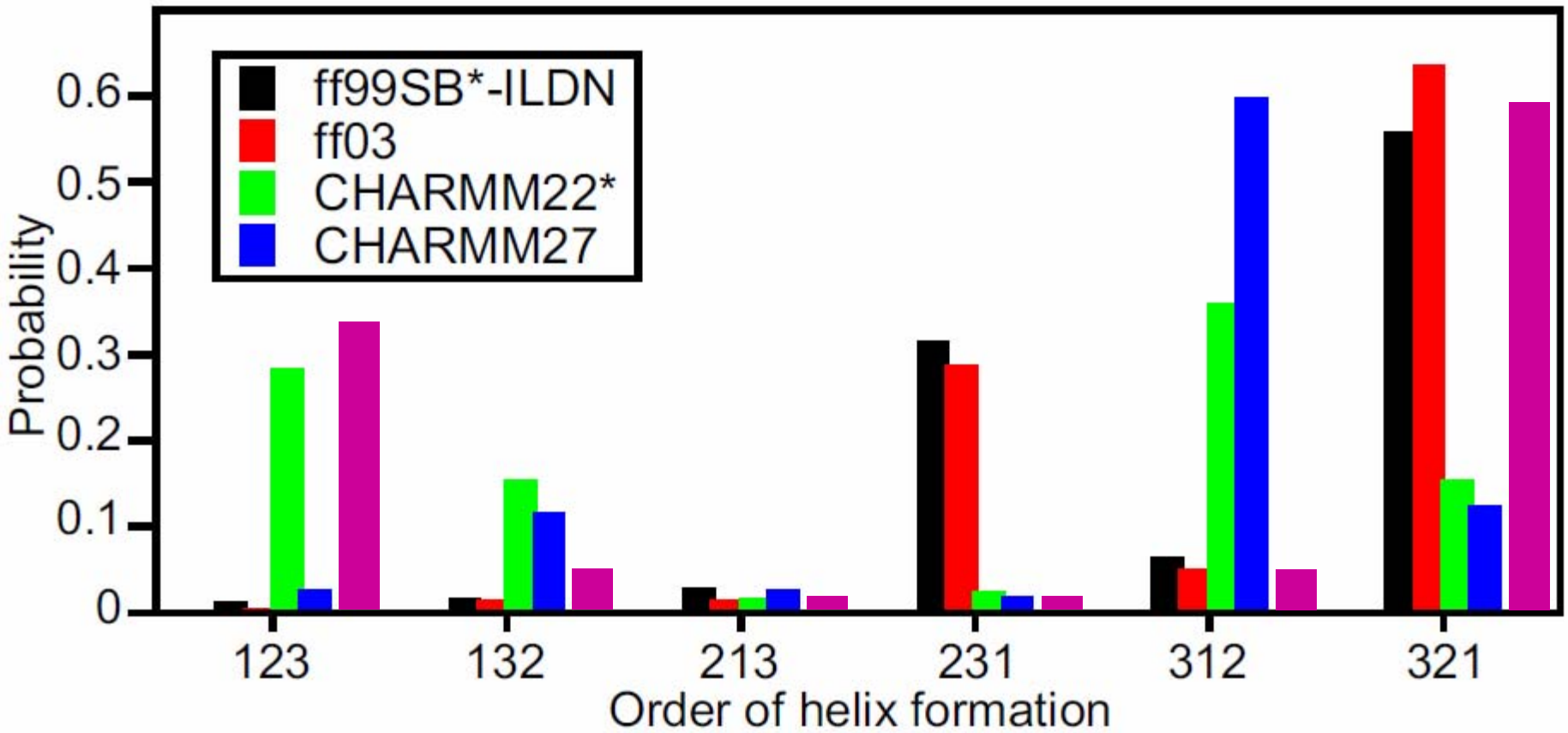
The transition path is defined as that portion of the trajectory after which the pfold is 0.2 and never falls below 0.2 before reaching the fully folded state (pfold = 1).

The order of helix formation is determined by the fraction of the transition path time during which the helix is formed.


1
0.5
0
pfold
0
100
200
300
step
transition
path


1
0.5
0
pfold
0
1000
2000
step
transition
path

Order of helix formation from transition paths of Muñoz-Henry-Eaton Ising-like model (for wild-type)



Why does such a simple model work so well?

Biology: evolved sequences have strong bias for forming native over non-native contacts (Wolynes)

Physics: coarse graining works because enthalpy-entropy compensation results in interaction free energies of ~ 1 ± 1 kcal/mol for all contacts (Fersht)

Implies a universal property:
overall fold determines both the rate and mechanism (?)

Coworkers in Laboratory of Chemical Physics, NIDDK, NIH

Experiments

Jan Kubelka
Troy Cellmer
James Hofrichter
Calorimetry: Jose Sanchez-Ruiz (U. Granada, Spain)

Theory:

Victor Munoz (1997 - 1999)
Eric Henry (2004 -.....)

Advice: Attila Szabo