UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL ATOMIC ENERGY AGENCY
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/916 - 29

SEVENTH COLLEGE ON BIOPHYSICS:

Structure and Function of Biopolymers: Experimental and Theoretical Techniques.

4 - 29 March 1996

Protein Folding - 1

Jose N. ONUCHIC
Department of Physics
University of California at San Diego
9500 Gilman Drive
CA 92093 La Jolla
U.S.A.

[redacted] SDSC 96

Towards a Realistic Folding Funnel for Protein

UCSD { JNO; N. Socci; M Betancourt; P. Leopold }

⊕ UI { PG Wolynes; Z. Luthey-Schulte }

Initial Funnel Idea

Numerical Simulation

Collaboration

Analytical ⊕ Numerical

Relation to Real systems

SUPPORT:

NSF

Beckman Foundation

OUR APPROACH —

HOW CAN WE UNDERSTAND THE FOLDING MECHANISM

- What are the key features of proteins that allow them to perform that?
- How do the proteins select the appropriate structure (out of 10^{60} for 60 amino acids in a relatively short time (Levinthal Paradox)?

BIOLOGICAL VIEWPOINT

- Is it something specific (unique) to proteins?

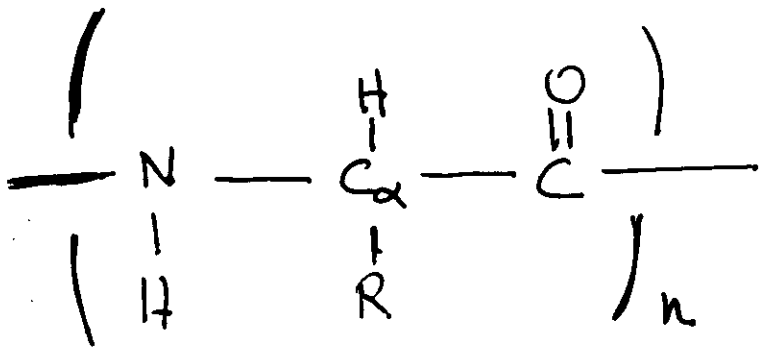
OR

PHYSICAL VIEWPOINT

- Are there generic, universal features shared by a class of systems which have these properties?

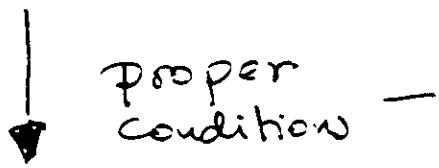
What are PROTEINS ?

Proteins are heteropolymers of Amino-Acids.

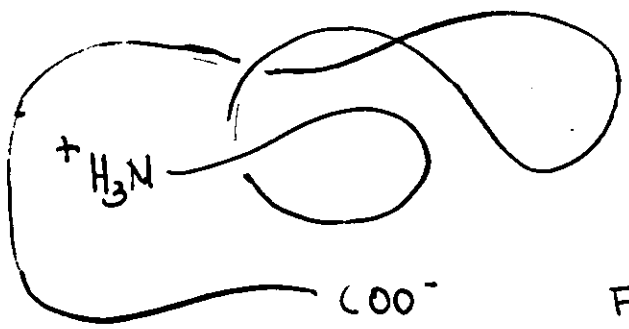


backbone

n — 40-400 small globular proteins



proper condition — physiological conditions



UNIQUE STATE

FOLDED STATE

STABILITY ~ 10 Kcal/mol (0.4eV)

- NOT ALL SEQUENCES FOLD!
  BIOLOGICAL ONES TO FOLD!

HOW ??

FOLDING MECHANISM

Anfinsen '60s

Small globular proteins (ribonuclease) fold (refold) in the absence of any catalytic molecule

↓

THERMODYNAMIC HYPOTHESIS
(Global Stable Minimum)

Levinthal '68

Too many states to probe to find the right one !!! PARADOX

Example: each A.A. 2 positions 2^N states $\begin{cases} N=100 \\ 10^{30} \end{cases}$

protein probes 10^{12} $\frac{\text{states}}{\text{sec}}$ ⇒ 10^{18} s

To save the T.H.

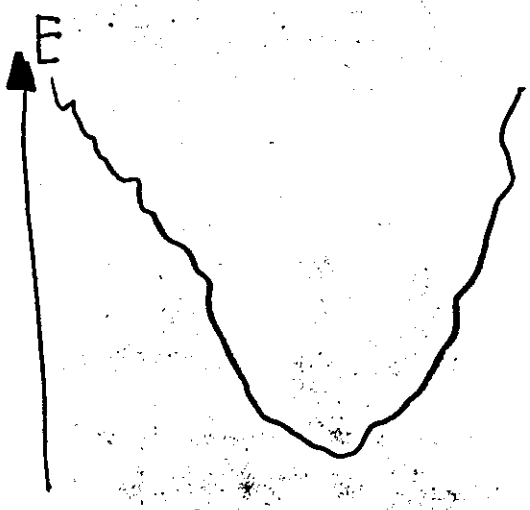
Baldwin '83 → Folding Pathway

REVISITING LEVINTHAL

BUT CRYSTALIZATION WORKS !!!

HOW ABOUT PROTEIN FOLDING ???

SMOOTH LANDSCAPE



configurational entropy loss is balanced energy gain from the slope

ROUGH LANDSCAPE

BUT REAL SYSTEMS HAVE SOME ROUGHNESS

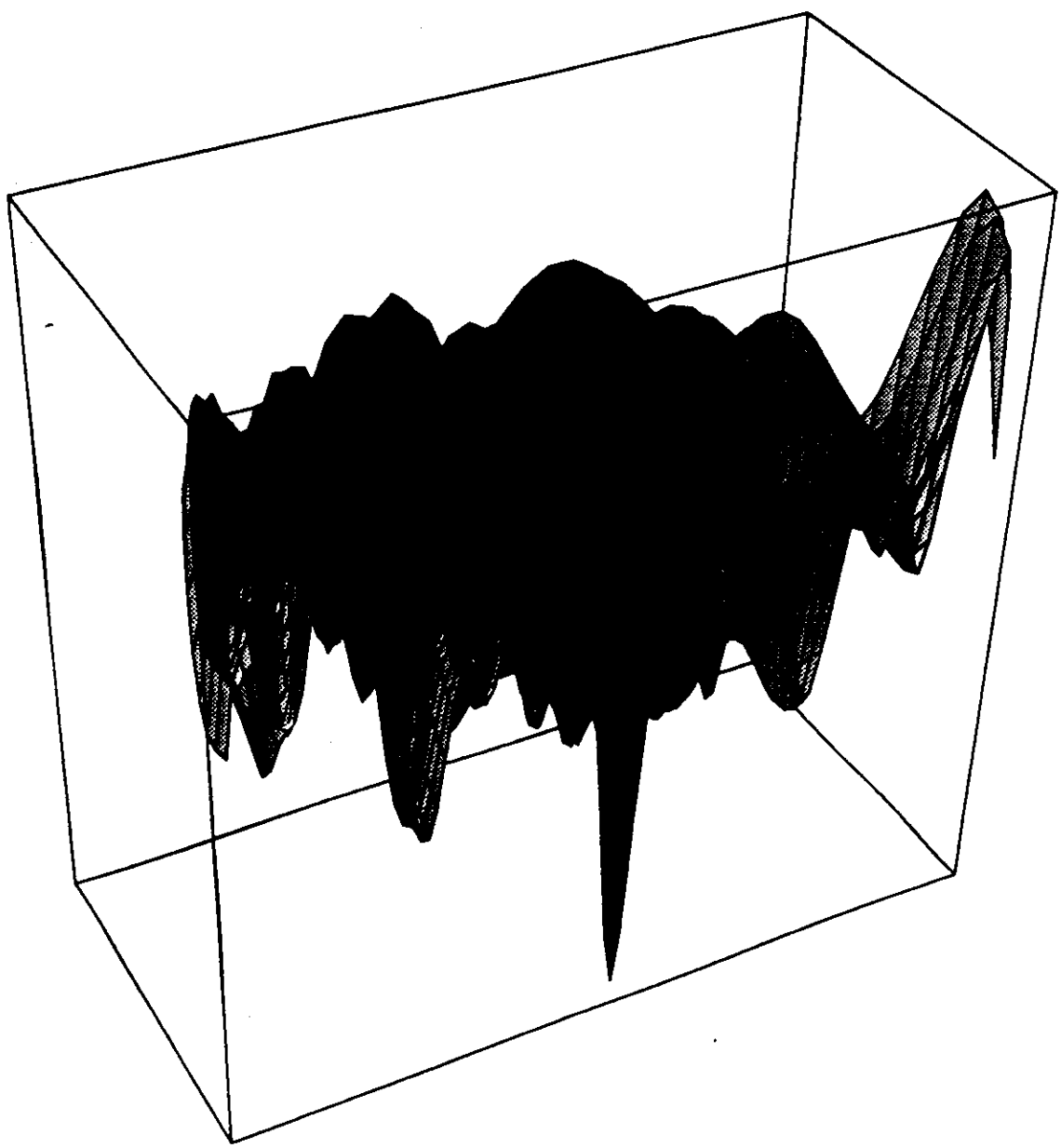
SOME FRUSTRATION

PRINCIPLE OF MINIMAL FRUSTRATION

Folding time considerable small compared to Levinthal

65 Wolynes

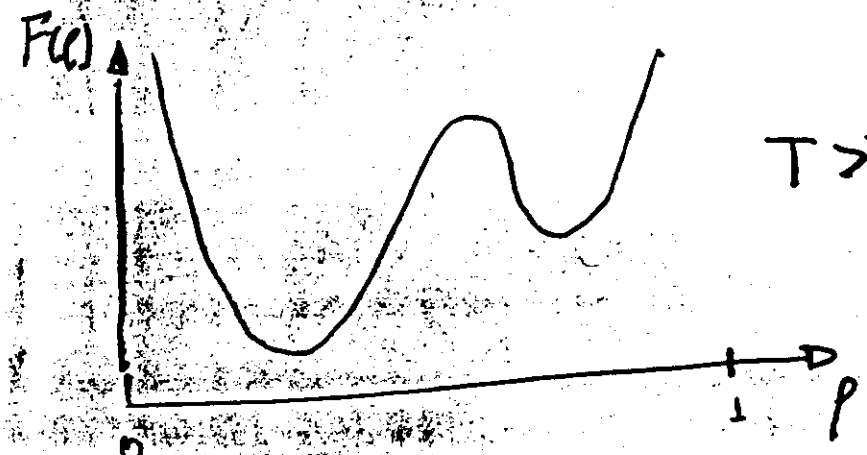
⇒ TO ADDRESS THIS QUESTION → SIMPLE MODEL



FOLDING THERMODYNAMICS

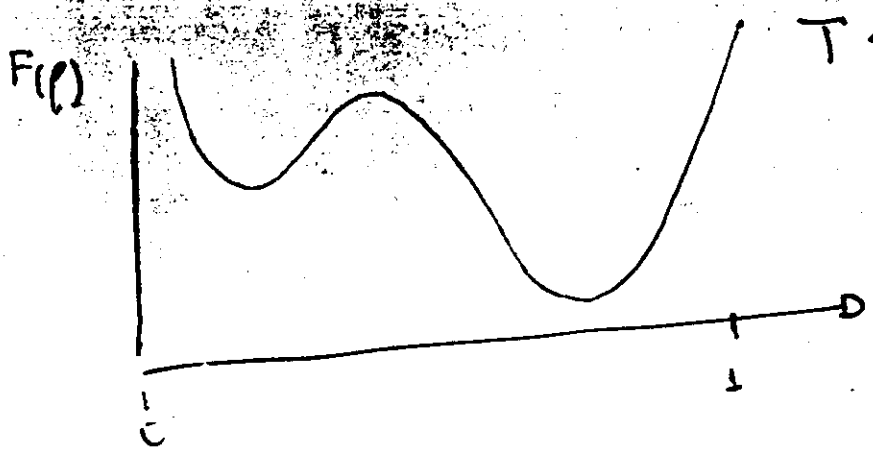
FOLDING IS AN "ALL OR NONE" TRANSITION

(FIRST ORDER LIKE)



$T > T_f$ "UNFOLDED"

ρ = % of folding



$T < T_f$ "FOLDED"

Physiological conditions

Double Well Free Energy Surface

SPIN GLASS IDEAS – Simple view

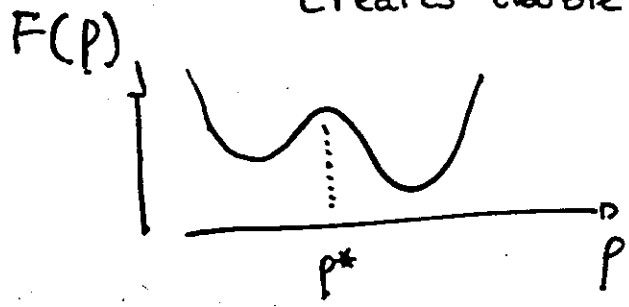
Bryngelson and Wolynes

Thirumalai
Wolynes
Bryngelson
Shakhnovich
Dill
Orland
⋮

(A) ORDER PARAMETER —CALL→ $\rho \equiv$ % of folding

ρ increases → $S(\rho)$ gets smaller
ρ increases → $\bar{E}$ gets smaller

Creates double well for $F(\rho)$



(B) Include roughness in the "potential" energy landscape

The "non-folded" part of the "protein" is treated as a Random Energy Model (REM)

* for every ρ there is a glass transition temperature $T_g(\rho)$

Bryngelson Wolynes Model — one order parameter ρ ⊕ REM | ENTROPY CRISIS TRANSITION

Good folding sequences → $T_\rho > T_g(\rho^*)$

Best Folding occurs $T_\rho > T > T_g$

Random Energy Model

Spin Glass

$$H_p(\{\sigma\}) = -\sum_{(i_1, i_2, \ldots i_p)} A_{i_1 \ldots i_p} \sigma_{i_1} \cdots \sigma_{i_p}$$

As p increases Energy configuration become random variables

$$\{\sigma_i^{(1)}\} \quad \{\sigma_i^{(2)}\} \qquad P(E_1, E_2) = P(E_1) P(E_2)$$

therefore

2^N energy levels

NJ^2 – variance of energy levels

$$P(E) = \frac{1}{\sqrt{N \pi J^2}} e^{-E^2/NJ^2}$$

if $n(E) \gg 1$

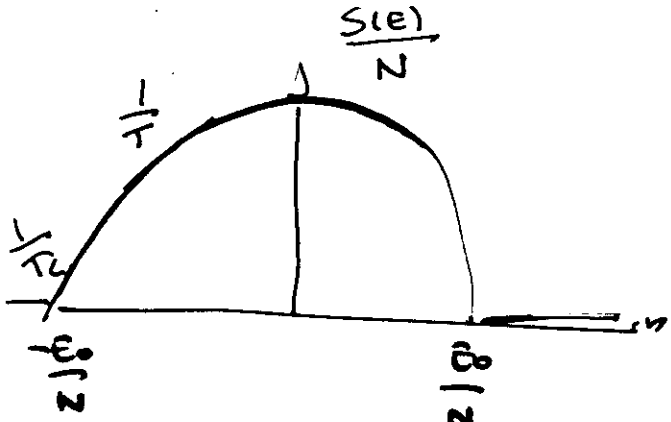
$$\langle n(E) \rangle = 2^N P(E) \sim \exp\left\{N\left[\ln 2 - \left(\frac{E}{NJ^2}\right)^2\right]\right\}$$

$S(E) \sim \ln n(E)$

$$= N\left[\ln 2 - \left(\frac{E}{NJ}\right)^2\right]$$

if $S(E) > 0 \Rightarrow |E| < E_0$

$$\frac{E_0}{N} = J\sqrt{\ln 2}$$



$$\frac{dS}{dE} = \frac{1}{T}$$

$S = 0 \Rightarrow \Omega = 2^N = e^{N S_c^*/k_B}$

$$T_c = \frac{J}{2\sqrt{\ln 2}}$$

$$k_B \ln \Omega = N S_c^* = \frac{N J^2}{(k_B T)^2}$$

S_c^* configurational entropy

$$\frac{F}{N} = -T \ln 2 - \frac{J^2}{4T} \qquad T > T_c$$

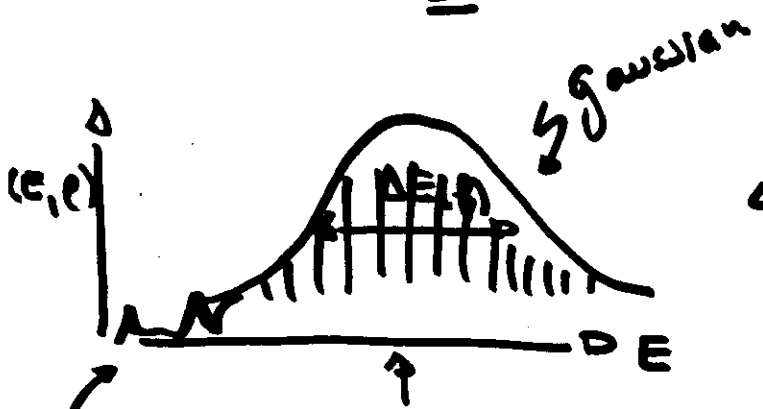
S.O.P.T

$$\frac{F}{N} = -\frac{E_0}{N} = -J\sqrt{\ln 2} \qquad T < T_c$$

A view on $T_g(\rho)$

In Principle
For every T there is a ρ
$T_g(\rho) = T$

For every ρ



Total $\Omega(\rho)$ states

deep states depend on sequence

mean depends on similarity to native state

$-E_{nat}\rho$

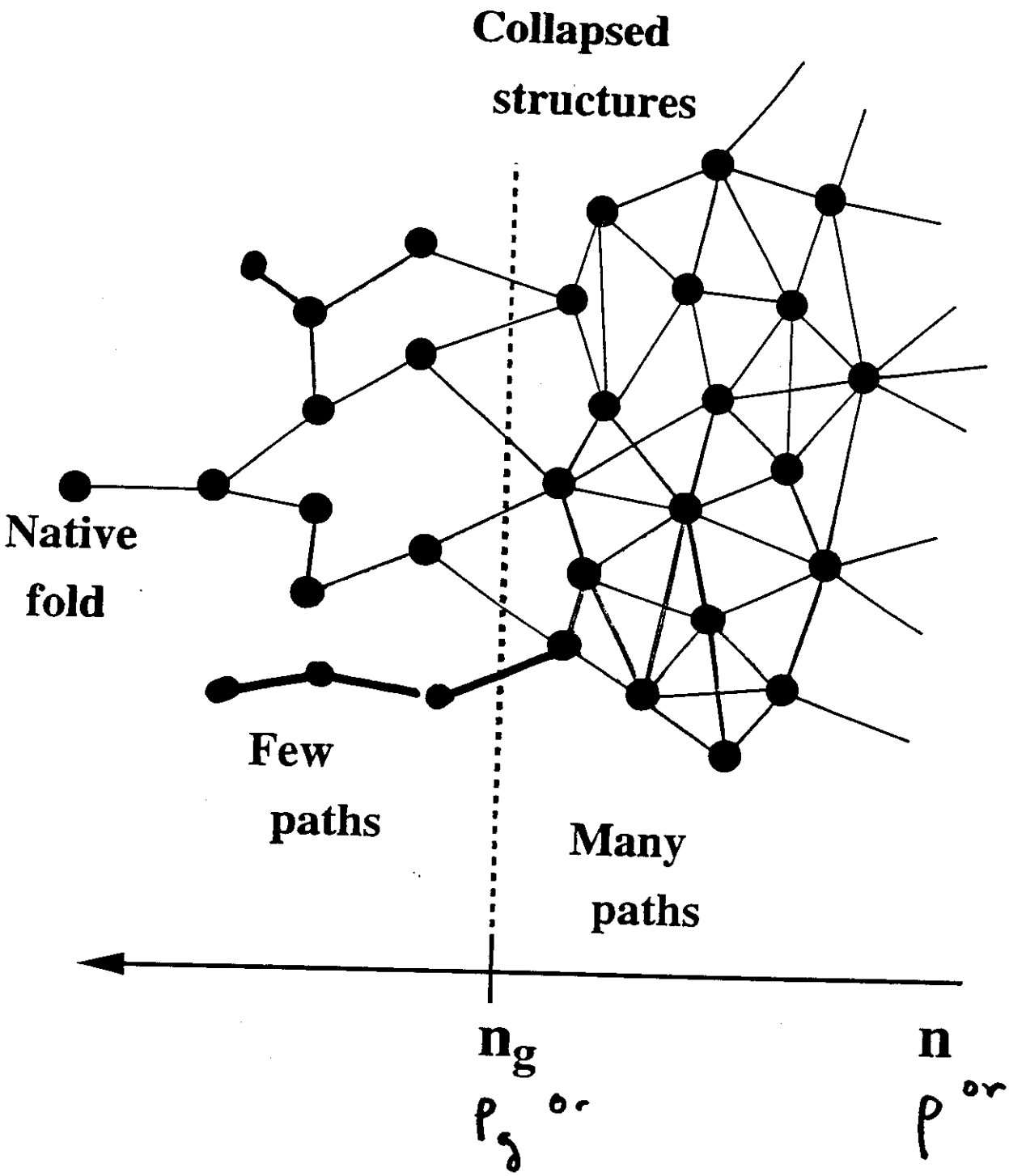
$\langle E \rangle = -E_{nat}\rho - \frac{\Delta E^2(\rho)}{2T}$

$S = k_B \ln \Omega(\rho) - \langle E \rangle^2 / 2\Delta E^2 \sim 0$

glass transition

$T_g(\rho) \sim \frac{\Delta E}{\sqrt{2 \ln \Omega(\rho)}}$

Figure 5


Collapsed
structures
Native
fold
Few
paths
Many
paths
n_g
ρ_g or
n
ρ or

Simple Models ⊕ Analytical Theories
is a powerful tool for studying
and understanding the
folding mechanism

LAW OF Corresponding States

Several other groups
working on simple models:

- Chan and Dill
- Thirumalai and Camacho
- Shakhnovich
- Orland
- Bryngelson
- Wolynes and Leviff
- .
- .
- .

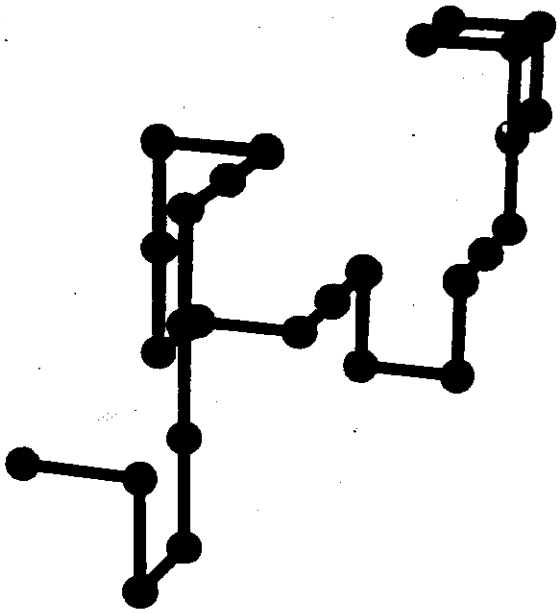
"More Realistic Models"

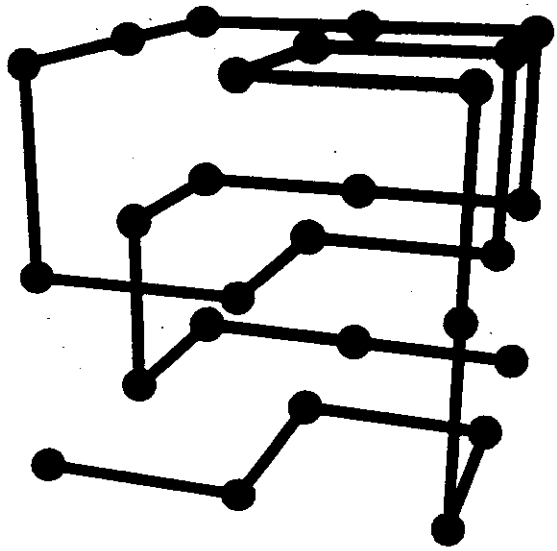
Jeff Skolnick

Charlie Brooks

ATOMISTIC
- Levitt and Daggett
- Scheraga
- Honig and Friesner

Model: 3D Simple Cubic Lattice Copolymer





- **27 monomers long.** **(Can not enumerate all conformations, but can enumerate cubes.)** N_c = MAX # CONTACTS 28
  Q = correct contacts

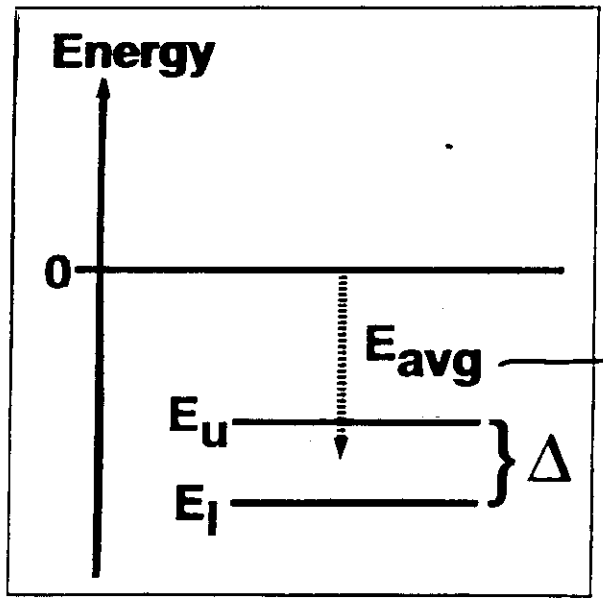
- Potential:

$$E = N_l E_l + N_u E_u$$

$$E_l < E_u < 0$$

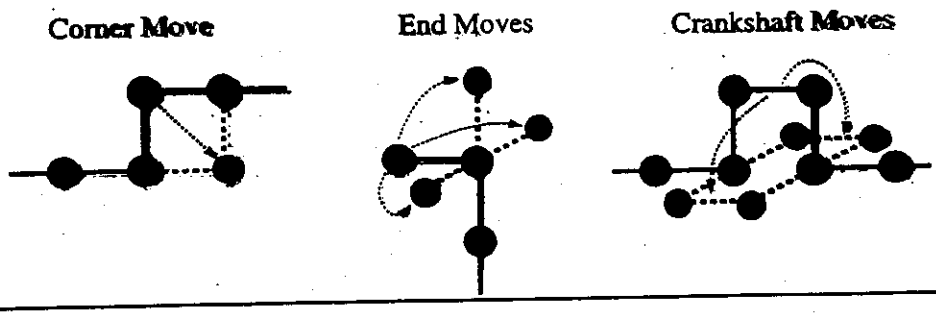
Initially

$$E_l = -3,\ E_u = -1$$



These values insure rapid collapse and that the ground state (native conformation) will be a cube, which we can find by enumeration

Dynamics—Local move Monte Carlo.[a]



Simulations

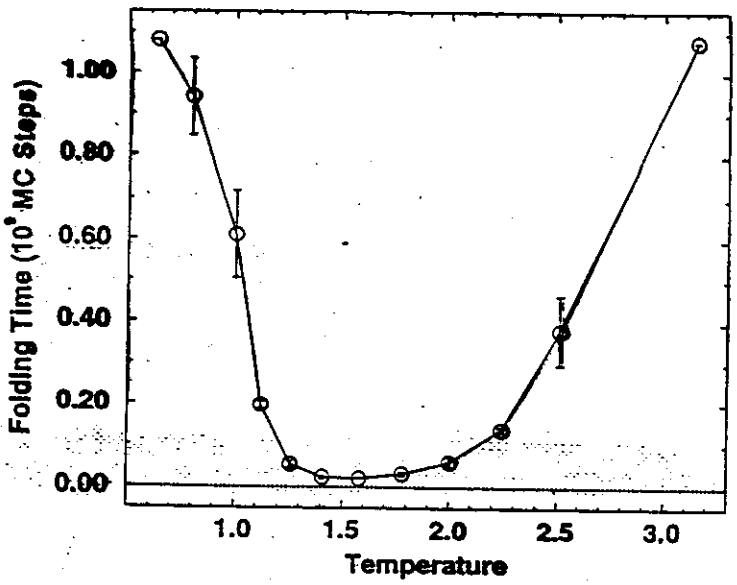
1. Kinetic
   - Starting from unfolded initial condition, run many folding simulations at various temperatures.
   - Maximum simulation time: $\tau_{max} = 1.08 \times 10^9$
   - Calculate mean first passage time (MFPT) for **collapse and folding**.
   - **Define kinetic glass transition** temperature: T_g
2. Thermodynamic—MC Histogram Method
   - Determine **folding temperature**, T_f. Compare with T_g for **several sequences**.

     Protein-like: $T_f > T_g$

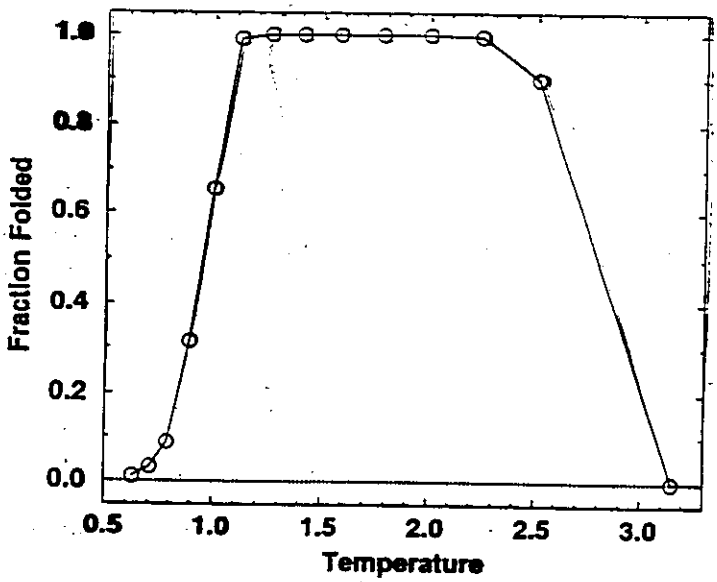
---

[a] SOCCI & ONUCHIC, *J.Chem.Phys.*, 101 1519 (1994)
J. Chem. Phys, in press

Kinetic Glass Transition



* Also in
2D - Chan, Dill, Haller
Asly - Sasai, Wang
Wolynes



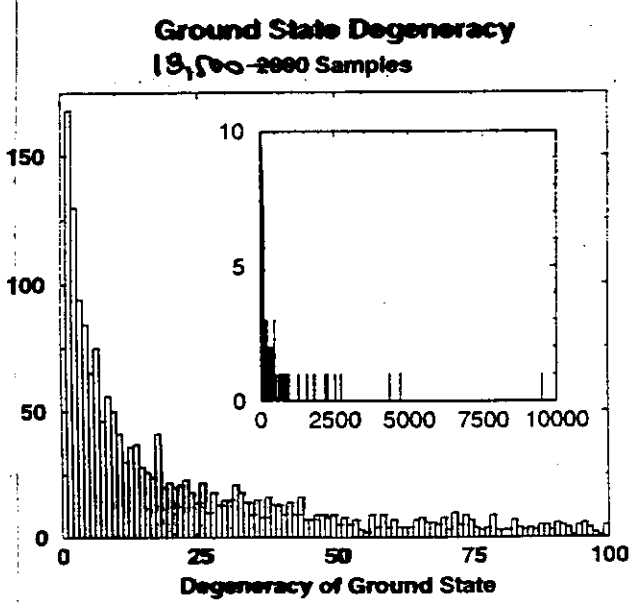
Temperature **at which** the system slows down (also loses **self-averaging** property). **Folding** time (τ_f) diverges **and** few runs find native state.

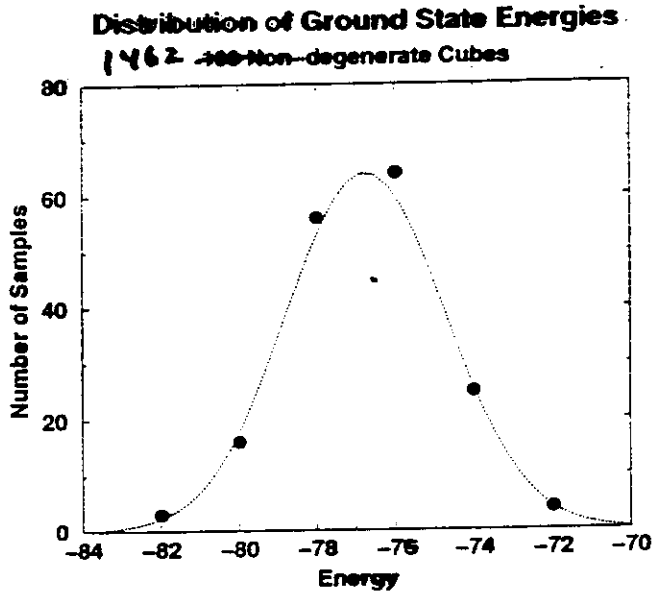
$$\tau_f(T_g) = \frac{1}{2}(\tau_{max} - \tau_{min}) \approx \frac{1}{2}\tau_{max}$$

$$T_g \approx 1$$

Run	Sequence	$E_{\min}$
002	ABABBBBBABBABABAAABBAAAAAAB	-84[a]
004	AABAABAABBABAAABABBABABABBB	-84
005	AABAABAABBABBAABABBABABABBB	-82
006	AABABBABAABBABAAAABABAABBBB	-80
007	ABBABBABABABAABABABABBBABAA	-80
013	ABBBABBABAABBBAAABBABAABABA	-76

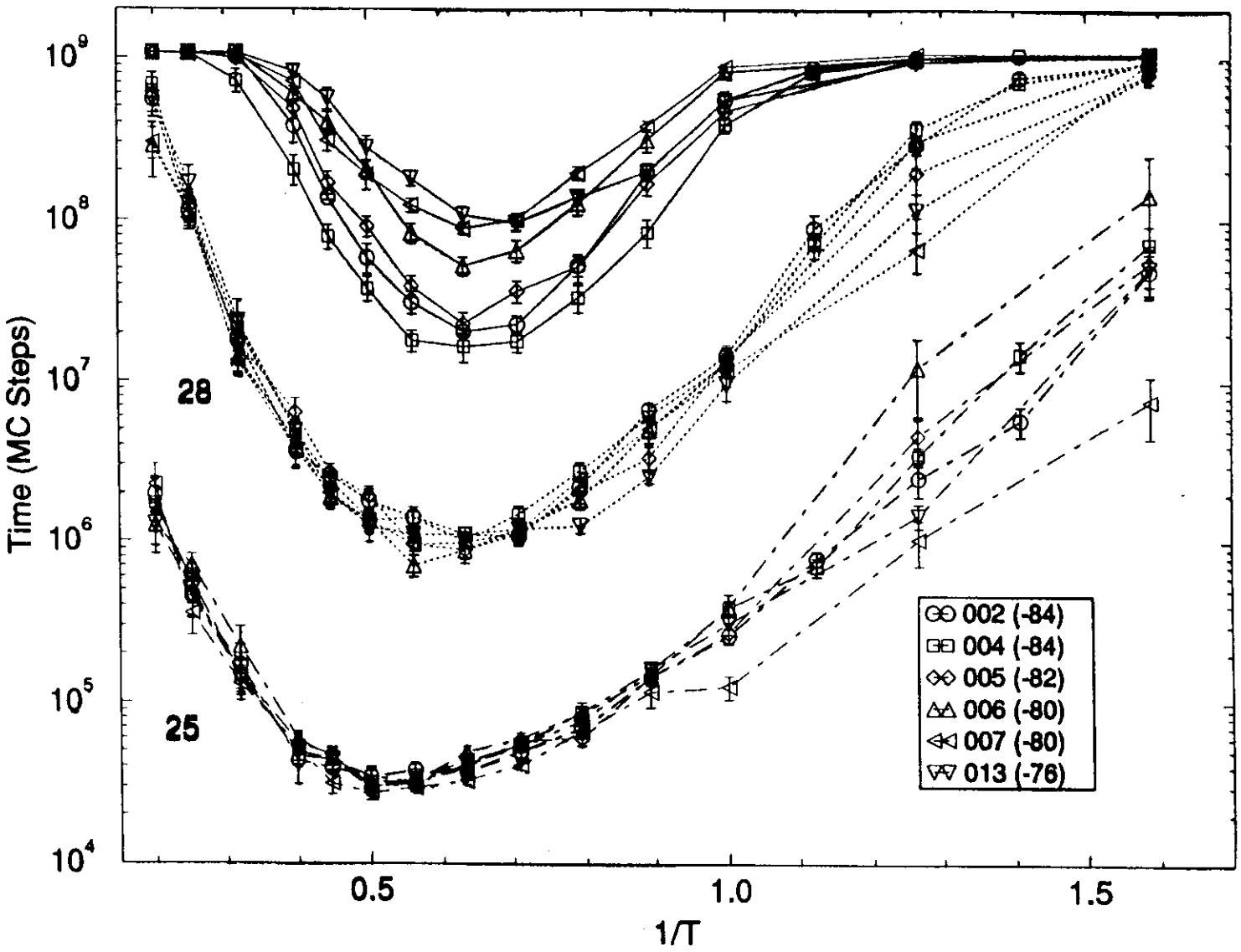
- All composed of 2 monomer types: same A:B ratio (14:13).
- All have unique ground states.
- Varying values for $E_{\min}$. Different amounts of "frustration," in ground state.
- 2 (-84) where designed, 4 choosen at random. (In 2000 samples did not get a -84.)





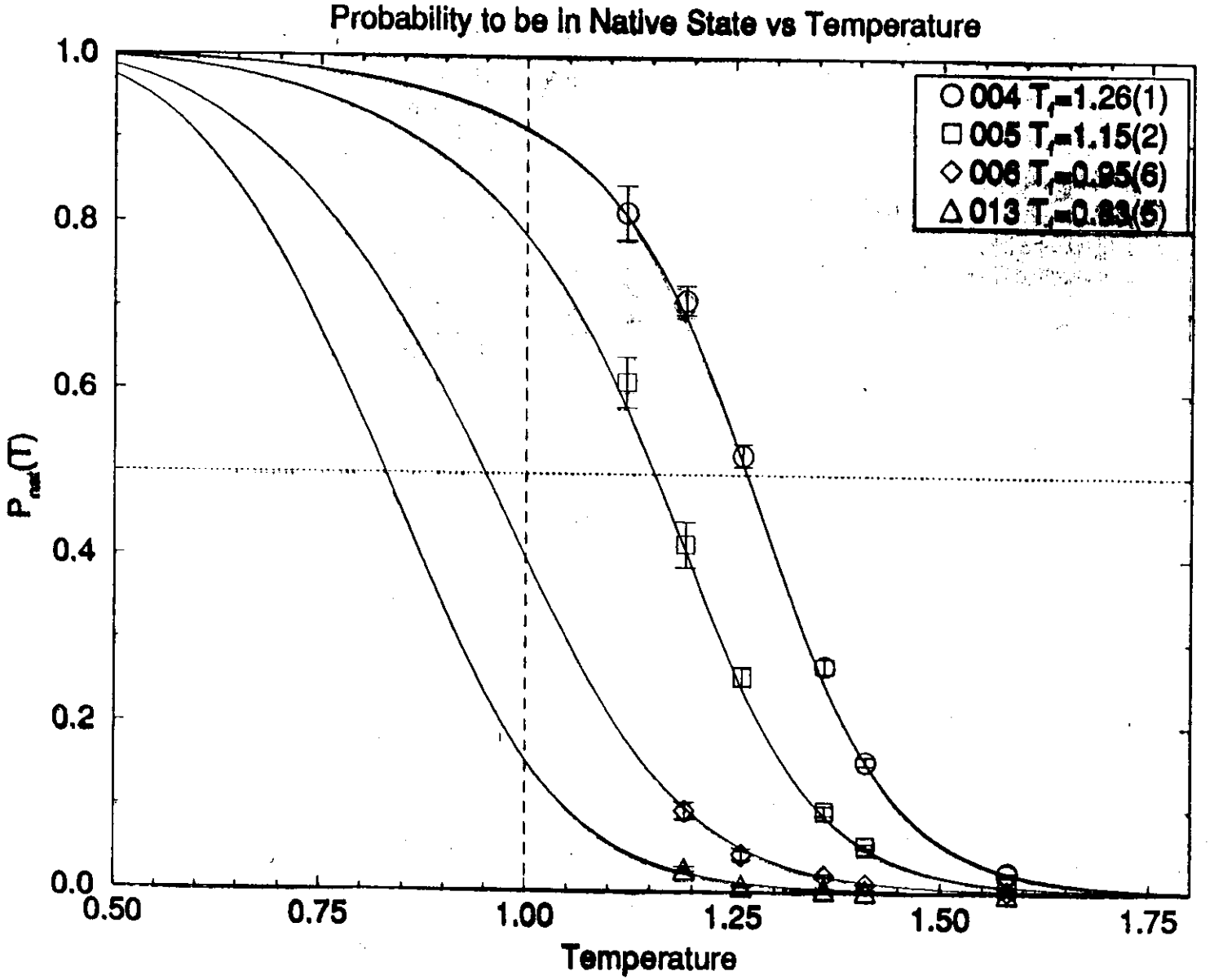
[a] Shakhnovich, *PNAS*, **90** 7195 (1993)

Folding and Collapse Times vs 1/T



Sep 18 20:10:08 1994

Folding Temperature



NDSocci

- 2 monomer type sequences (*2 LC*) are qualitatively protein-like ($T_f > T_g$), but not too good quantitatively.

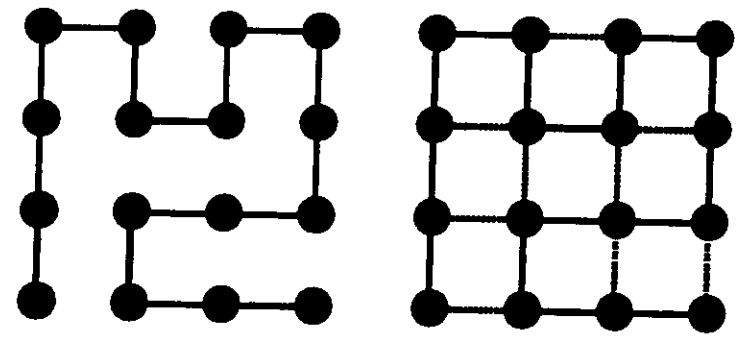
  $T_f/T_g \approx 1.3$ is a little too small.

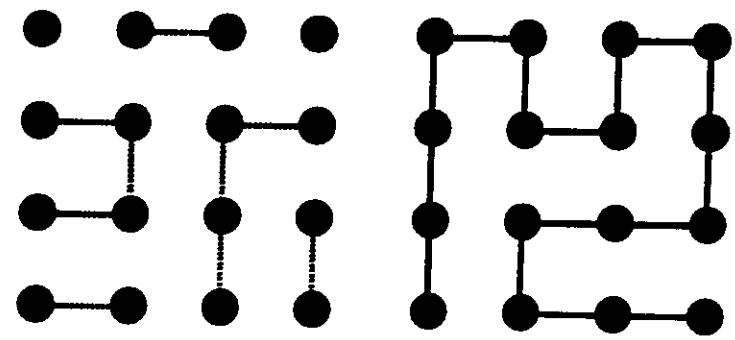
- Can we tune the model to give a higher T_f/T_g ratio? Ie, can we make it more protein like?

- Change number of monomer types (*2→3 LC*).

  Idea: reduce "energetic frustration" in the collapsed states by reducing the number of strong by incorrect (non-native contacts).

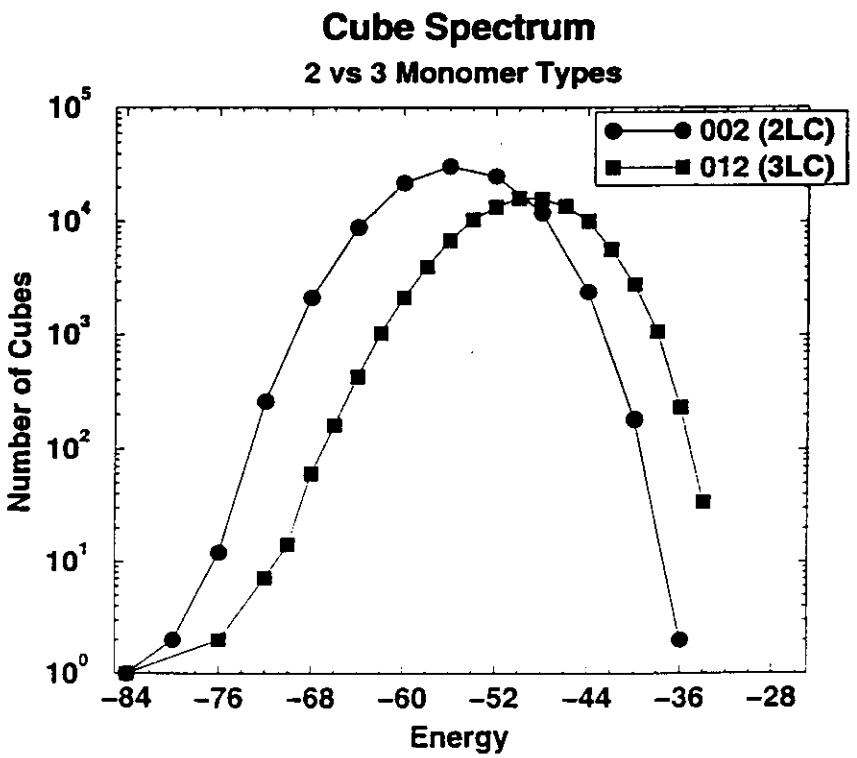
- Design sequences by domain decomposition.

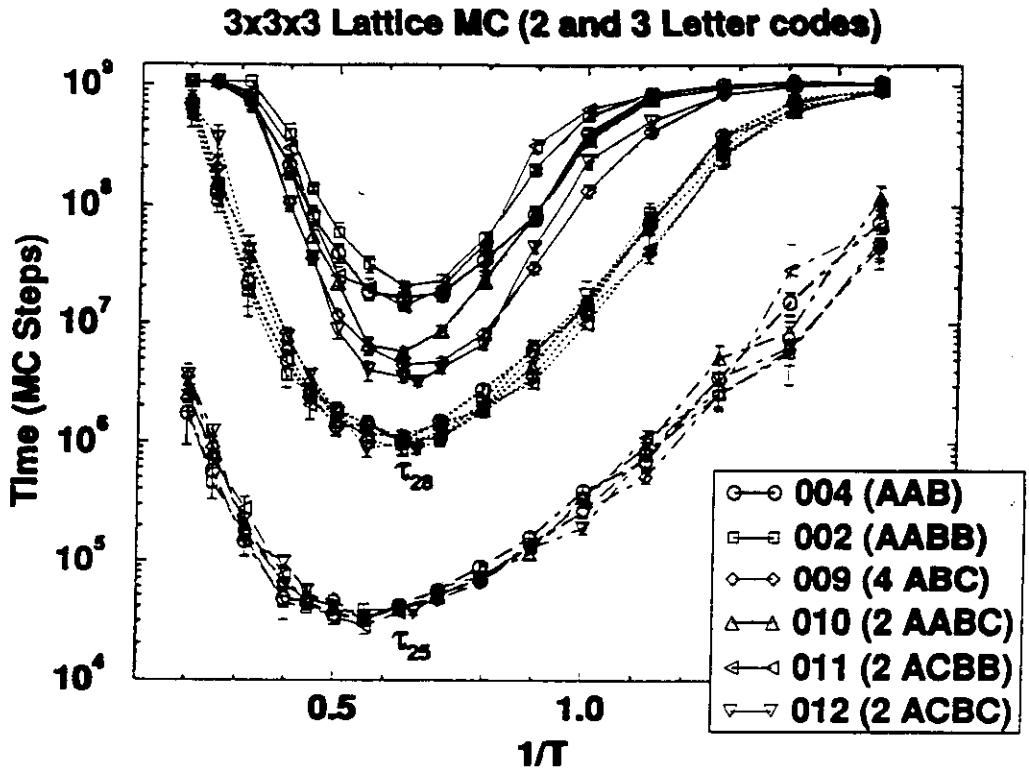


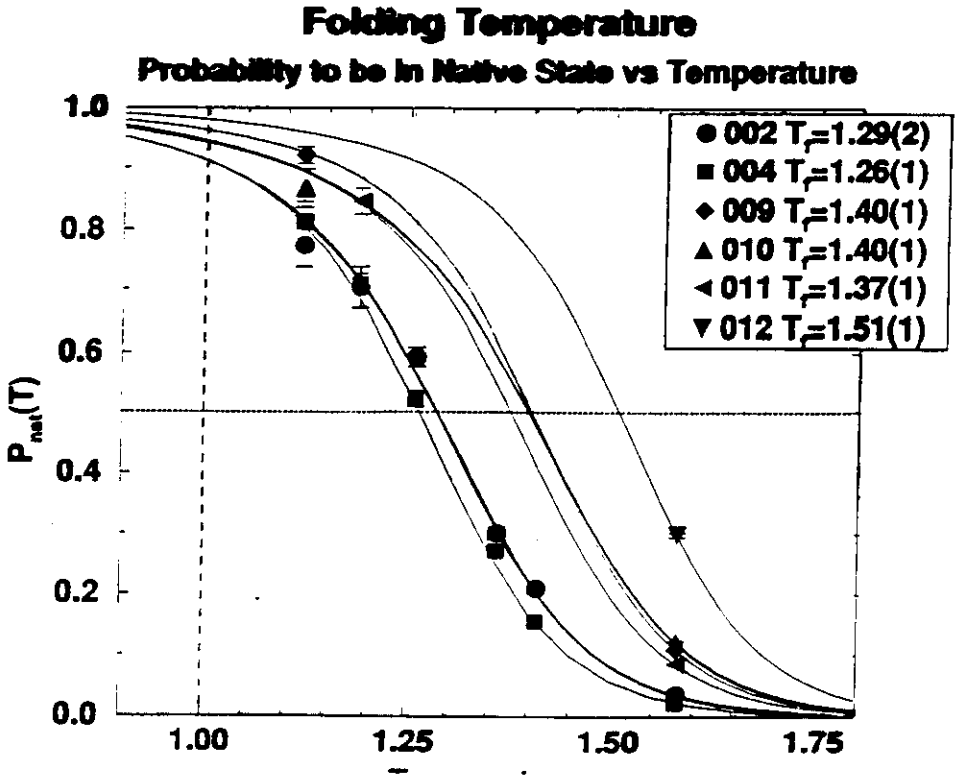


Run	Sequence	τ_{min}	T_g	T_f
002	**ABABBBBBABBABABAAABBAAAAAAB**	2.0×10^7	1.02	1.285(15)
010	**ababbbCbaCbababaaaCbaaaaaab**	6.2×10^6	0.94	1.40(1)
011	**ababbbabaababababaCaabaCaaCab**	1.4×10^7	1.03	1.37(1)
012	ababbbCbaCbababaCaCbaCaaCab	2.9×10^6	0.91	1.509(1)
004	AABAABAABBABAAABABBABABABBB	1.6×10^7	0.99	1.26(1)
009	aaCaabaabbabaaaCabCababab**bb**	4.3×10^6	0.86	1.40(1)

- Energy of native (ground state) remains unchanged.
- Energies of many compact non-native states are increased.
- Smoothes out the energy landscape, removing traps.




3x3x3 Lattice MC (2 and 3 Letter codes)
Time (MC Steps)
1/T
τ_{26}
τ_{25}
004 (AAB)
002 (AABB)
009 (4 ABC)
010 (2 AABC)
011 (2 ACBB)
012 (2 ACBC)


Folding Temperature
Probability to be in Native State vs Temperature
$P_{nat}(T)$
002 T_f=1.29(2)
004 T_f=1.26(1)
009 T_f=1.40(1)
010 T_f=1.40(1)
011 T_f=1.37(1)
012 T_f=1.51(1)

- Adding a third amino acid type, can decrease folding time by almost a factor of 5 (collapse and folding converge).

- Can lower T_g slightly and increase T_f dramatically.

$$T_f/T_g \to 1.6 \quad (\approx 20\%)$$

- Concept of *Minimal Frustration*. As energy decreases conformations how greater overlap with the native state. Truer for the 3LC sequences than the 2LC.

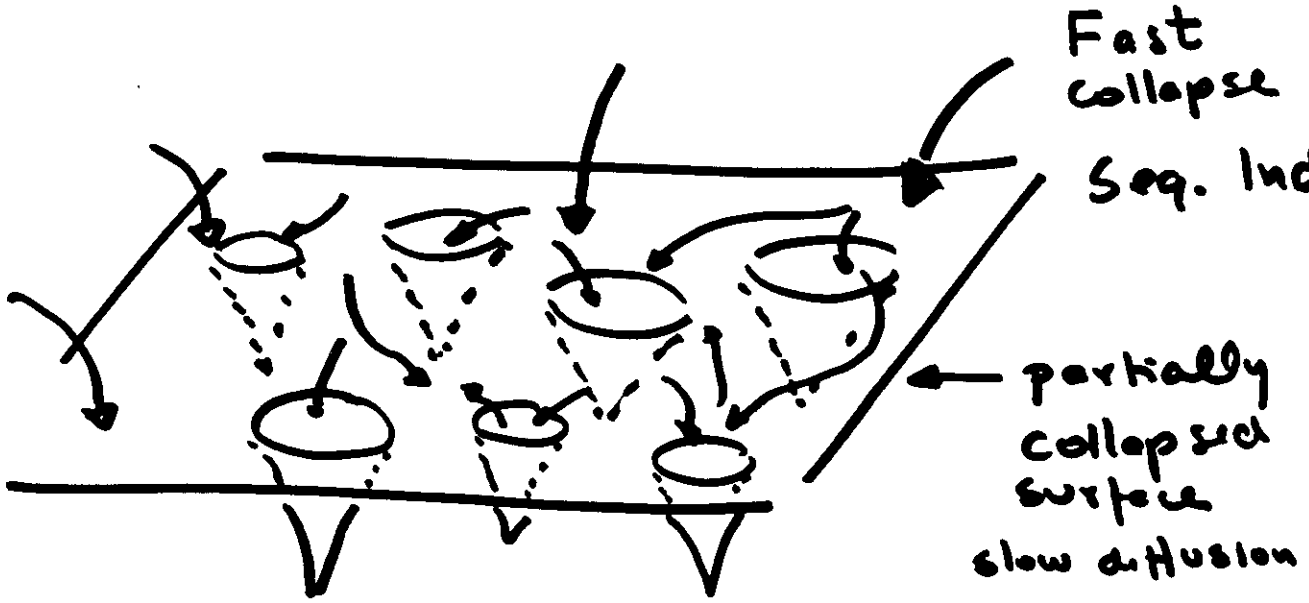
- Using a 3LC sequences combined with analytical models for chain collapse, we are beginning to make connections with experimental data for real proteins.[a]

---

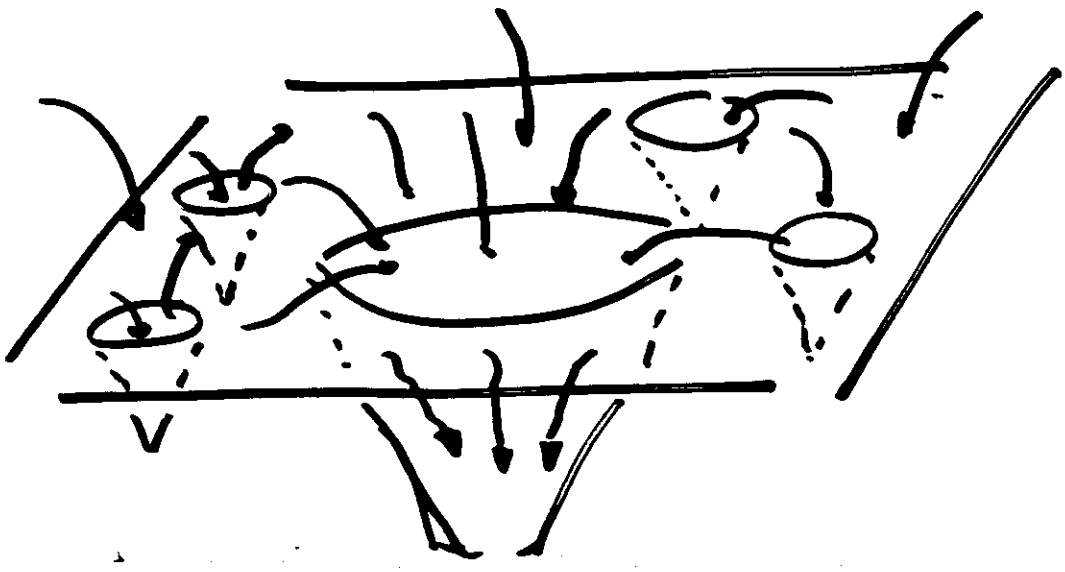
[a] ONUCHIC, WOLYNES, LUTHEY-SCHULTEN & SOCCI, *PNAS*, in press.

What do we learn from simple models

- Ruggedness large compared to stability
  Random Heteropolymer



- Ruggedness small compared to stability
  Protein Sequences



NEW FOLDING PICTURE

Onuchic
Wolynes
Luthey-Schu
Socci

- Proteins Have "2" Energy Scales
  - H-Bonds
  - Interactions Non-Local Sequence but Local space !!

LAW OF CORRESPONDING STATES

(1) EXAMPLE – Small α-Helical Protein

60 AA protein

Collapsed Molten Globule State

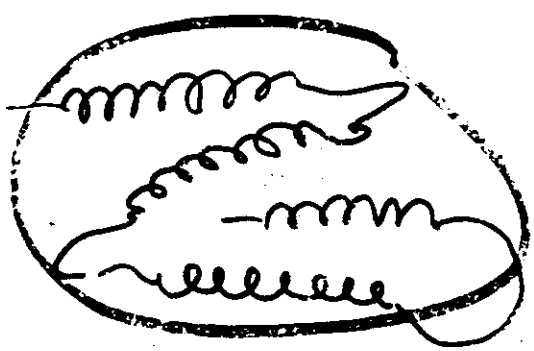
(High Degree of Helicity ↔ Not necessarily native helics)

27-mer lattice model

Similar Levinthal entropy

"Lattice Model"

Ramirez
Schulten
Wolynes



Exp. Support

Callender + Los Alamos Dyer et al

Kalleubach - Apomyoglobin

Engelman - Staphylococcal nuclease

Kinetic View

Reconfiguring among these states

3 IMPORTANT PARAMETERS

1. Size of Landscape ≡ Levinthal Entropy δ_L
2. Ruggedness of Landscape · $\frac{T_f}{\Delta E^2}$ MAX
3. Slope of Funnel → stability gap · $\frac{T_f}{\delta E_s}$

(II) Models with Similar Parameters → Similar Mechanism

LAW OF CORRESPONDING STATES

relating real proteins with simple model

(III) Need of experiments to study "early folding" events → Bill Eaton !!! ...

$\frac{T_f}{T_g}$ Real Proteins

Comparison of motions for free chains and collapsed chains

$$\tau_{reconfg} = \tau_{free}\, e^{\frac{\Delta E^2}{2T^2}}$$

$\tau_{free} \sim \frac{n^3}{}$ { simulations, Rouse-Zimm theory

$\tau_{reconfg} \sim ms\text{-}s$ { NMR e.g. lactalbumin, Baum; apo-cytochrome b562, Wand

$$\frac{\Delta E^2}{2T_f^2} = 11\text{-}18$$

Folding Transition

$$\frac{\delta E_S}{T_f} = S_L + \frac{\Delta E^2}{2T_f^2}$$

energy loss down the funnel compensate entropy loss

$$T_g^2 \frac{\Delta E}{\sqrt{2S_L}} \Longrightarrow \frac{T_f}{T_g} \sim \underline{1.6}$$

$S_L \sim 30\text{-}40$

Order Parameters

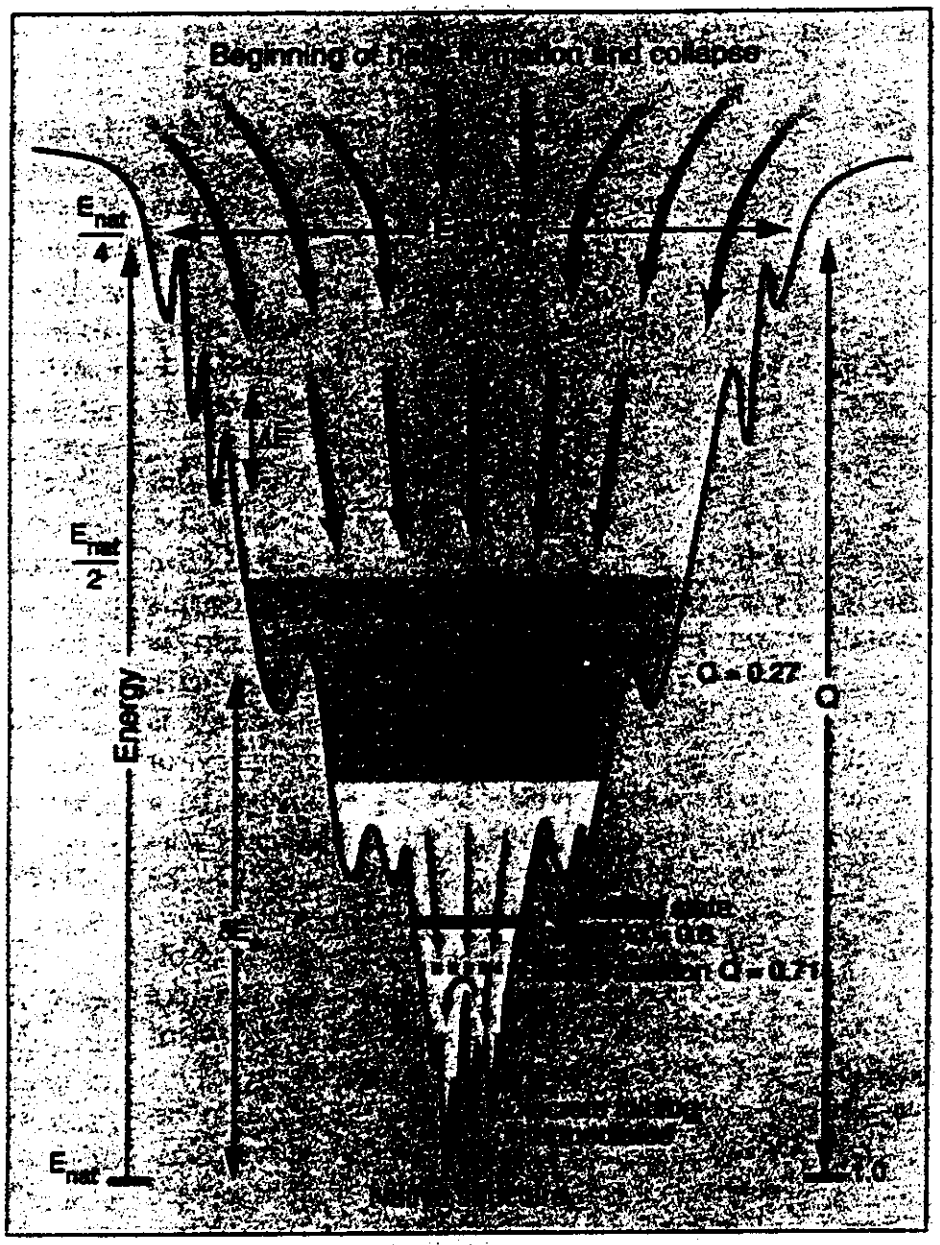
N_c = # contacts 0-23 collapse

Q = # native contacts 0-23

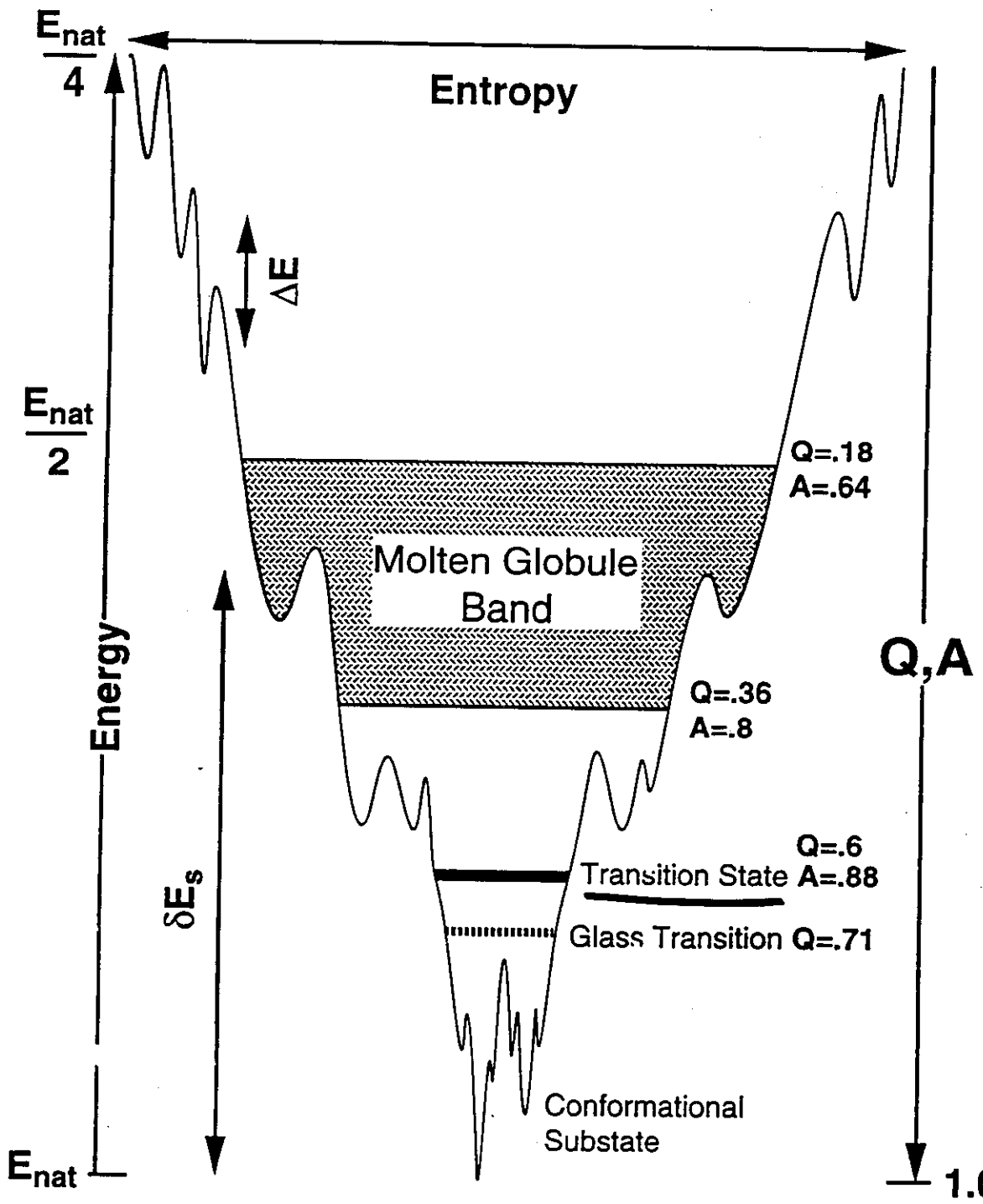
A = # native angles 0-25

Bryngelson-Wolynes

$E_{NATIVE} : +28 \times (-3) = -84$


Beginning of
$\frac{E_{nat}}{4}$
$\frac{E_{nat}}{2}$
Energy
Q = 0.27
Q
Q = 0.71
E_{nat}
1.0

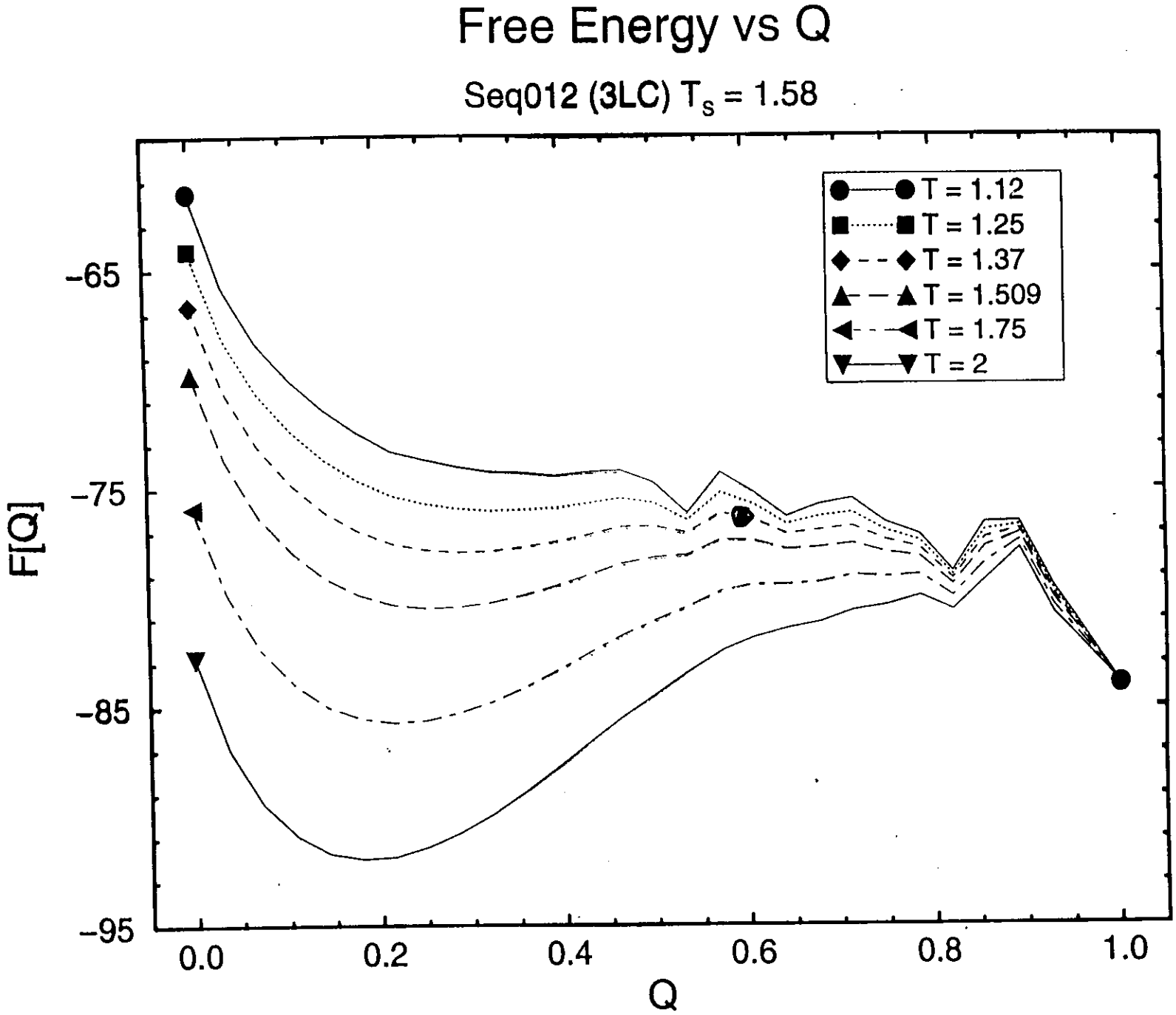
Q = 0.6 in TS okay Fsubt (CI2)



$\frac{\delta E_s}{\Delta E} \sim 1.4$

$\frac{T_f}{T_g} \sim 1.6$

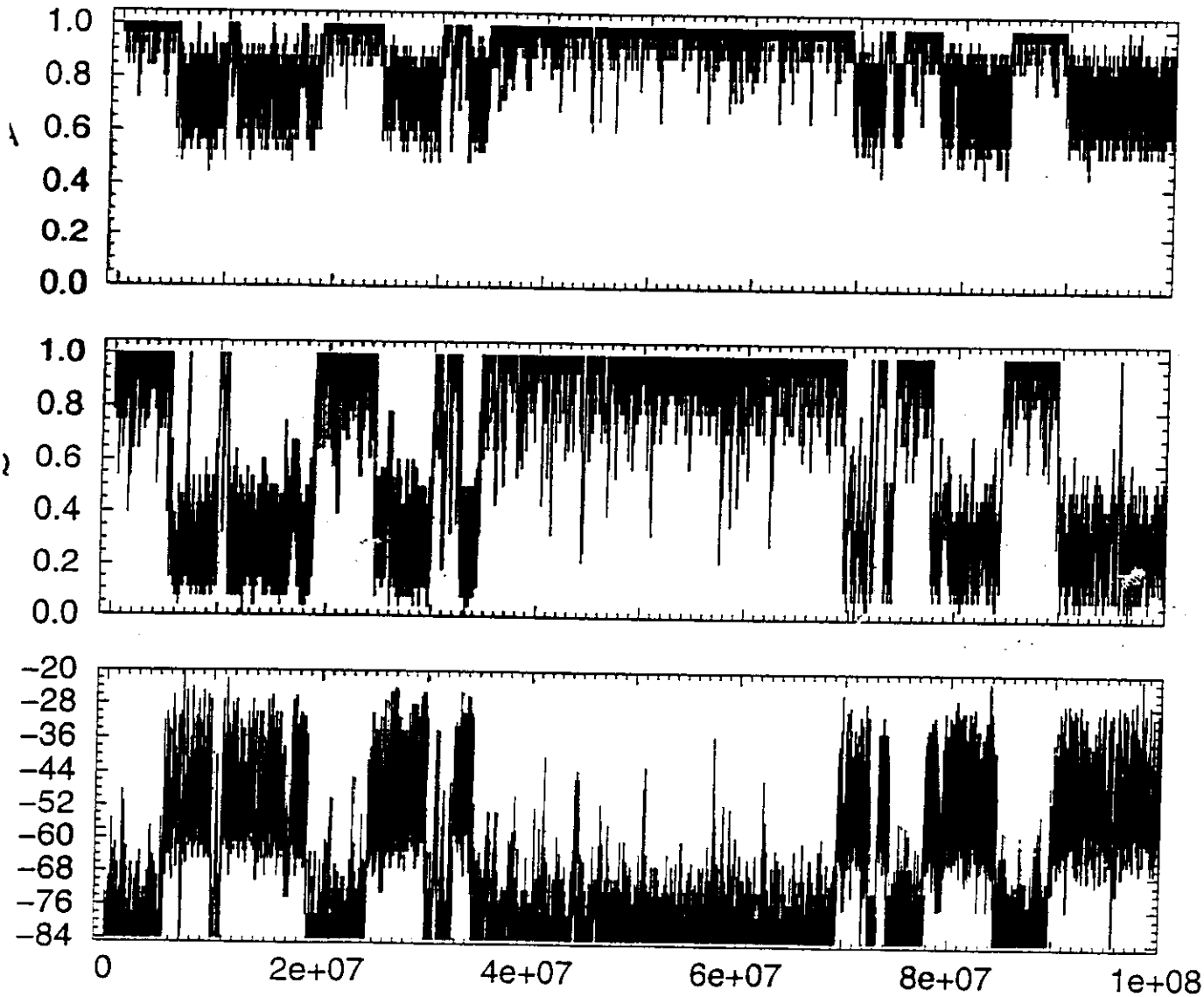
Based on correspondence
stats to 3-letter
code

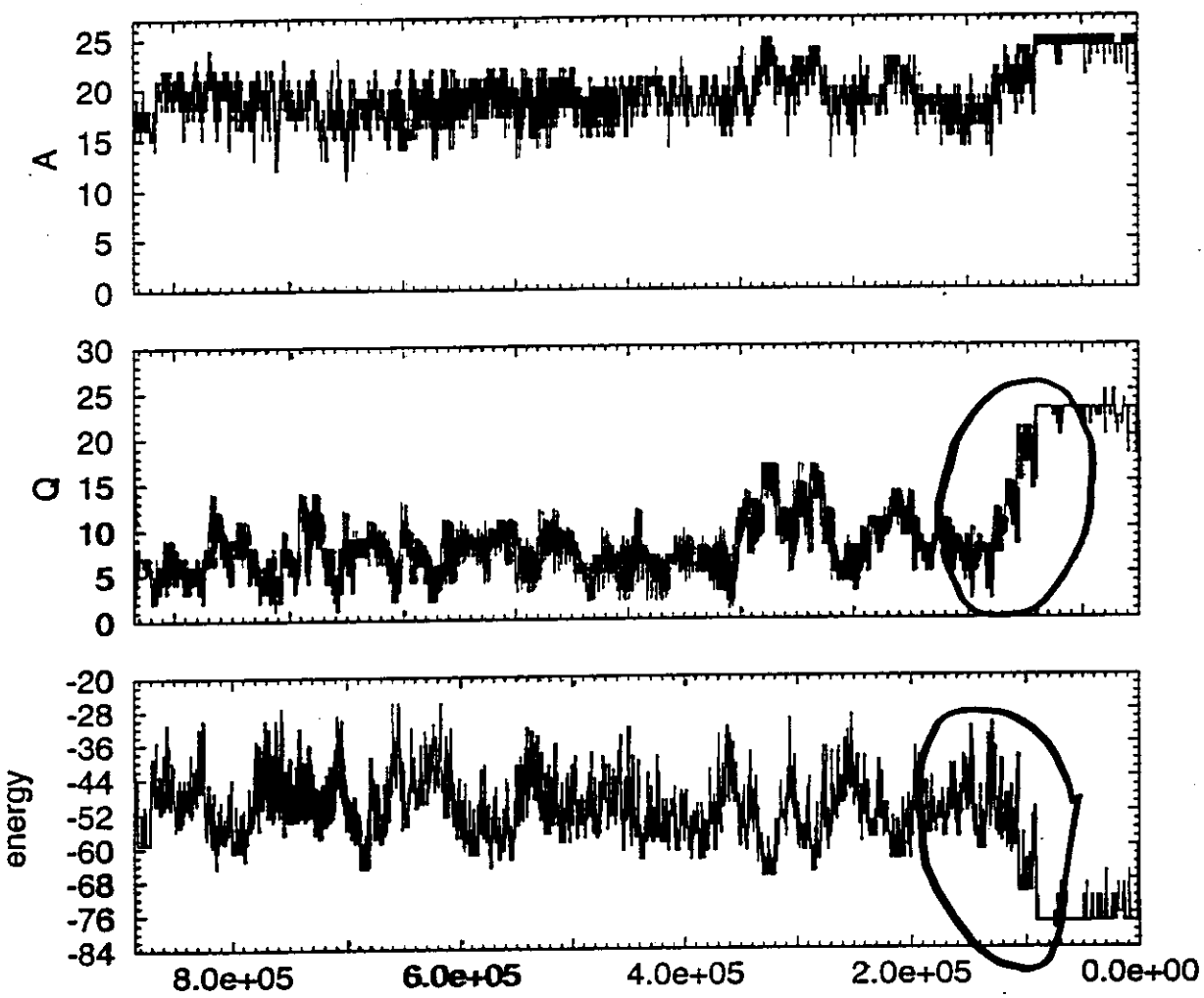


Thu Aug 10 22:39:18 1995

NDSocci

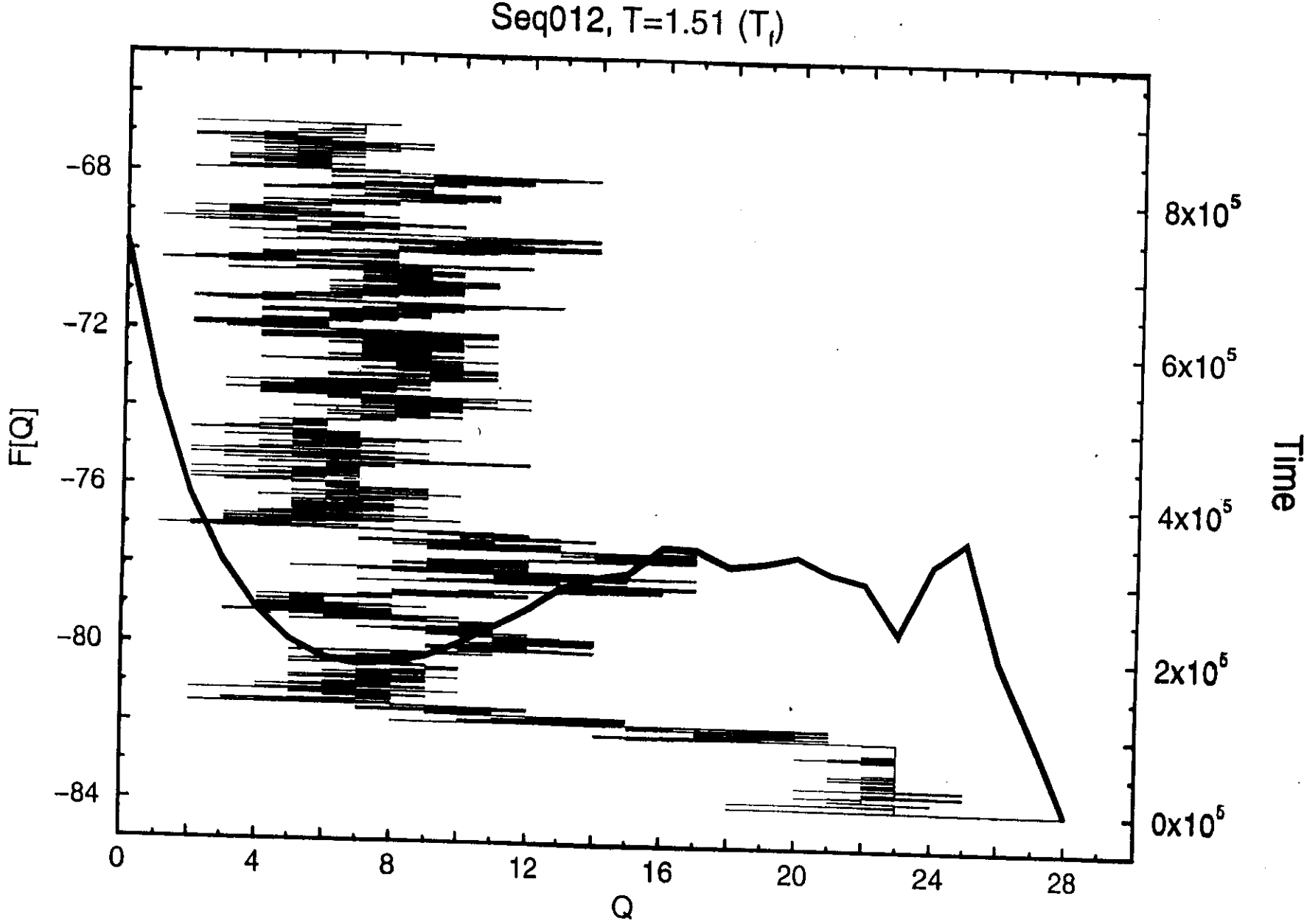
$\tau_f \sim 3 \times 10^6$





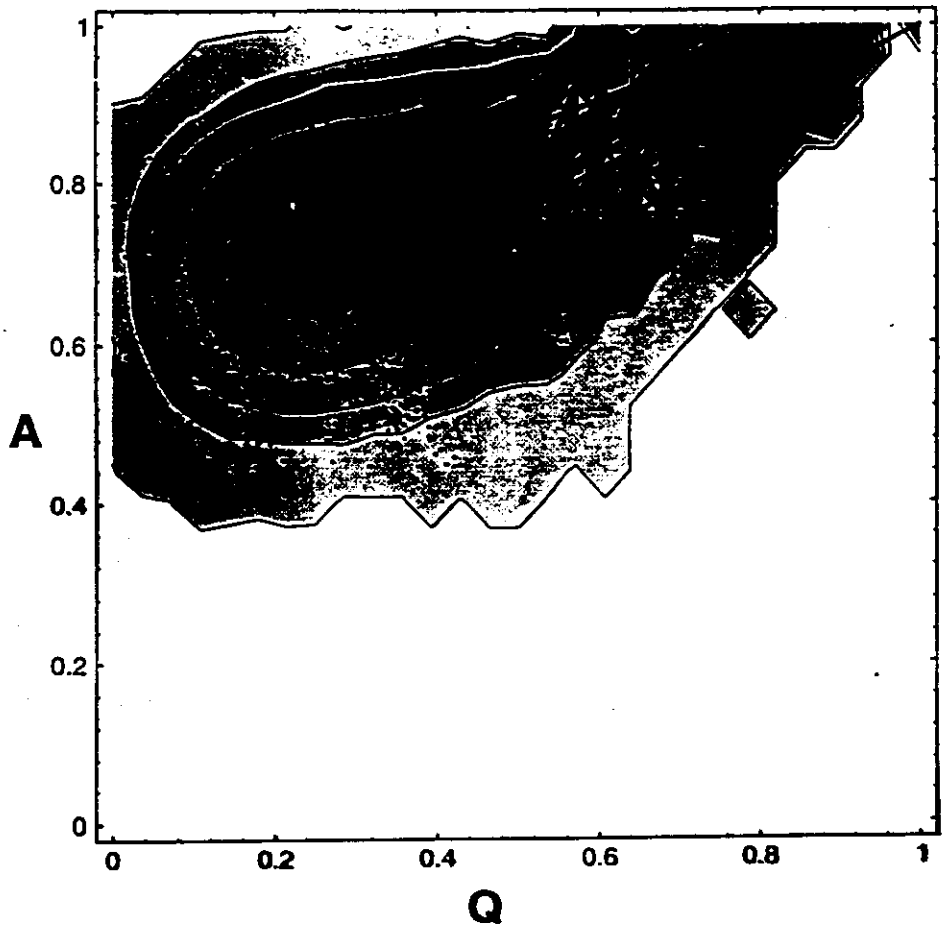
~100,000 steps

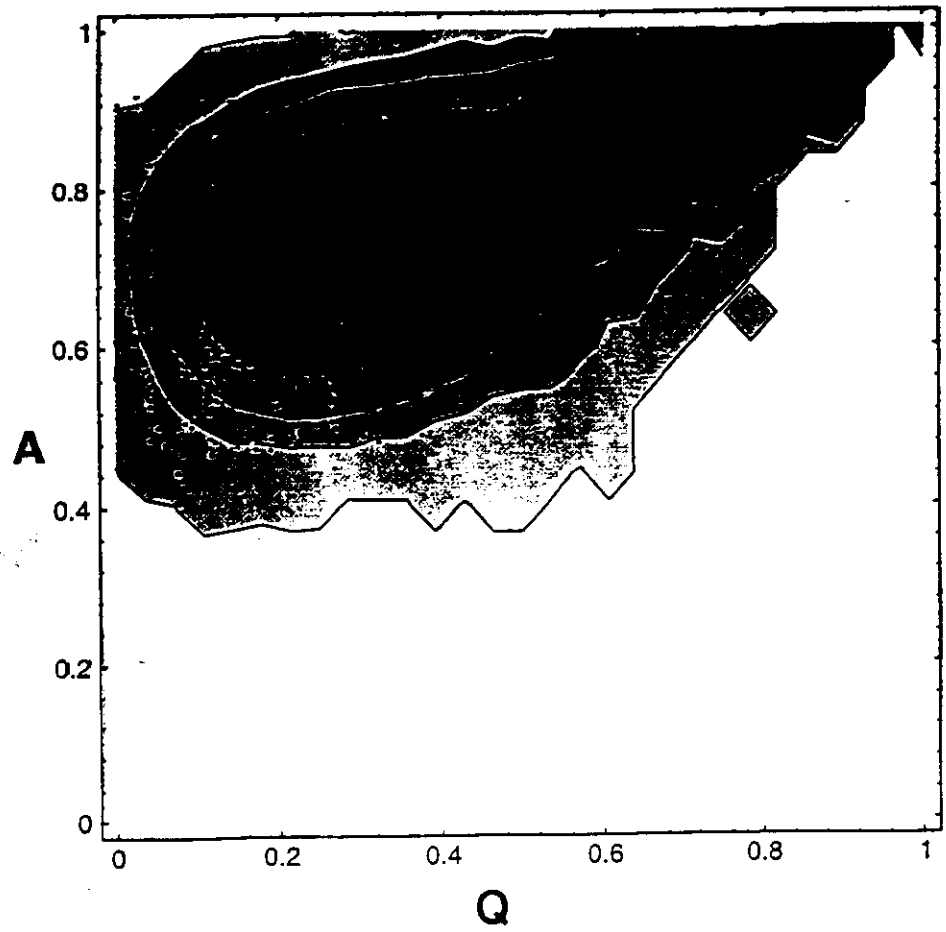
Free Energy and Folding Traj vs Q



Mon Oct 16 21:43:08 1995

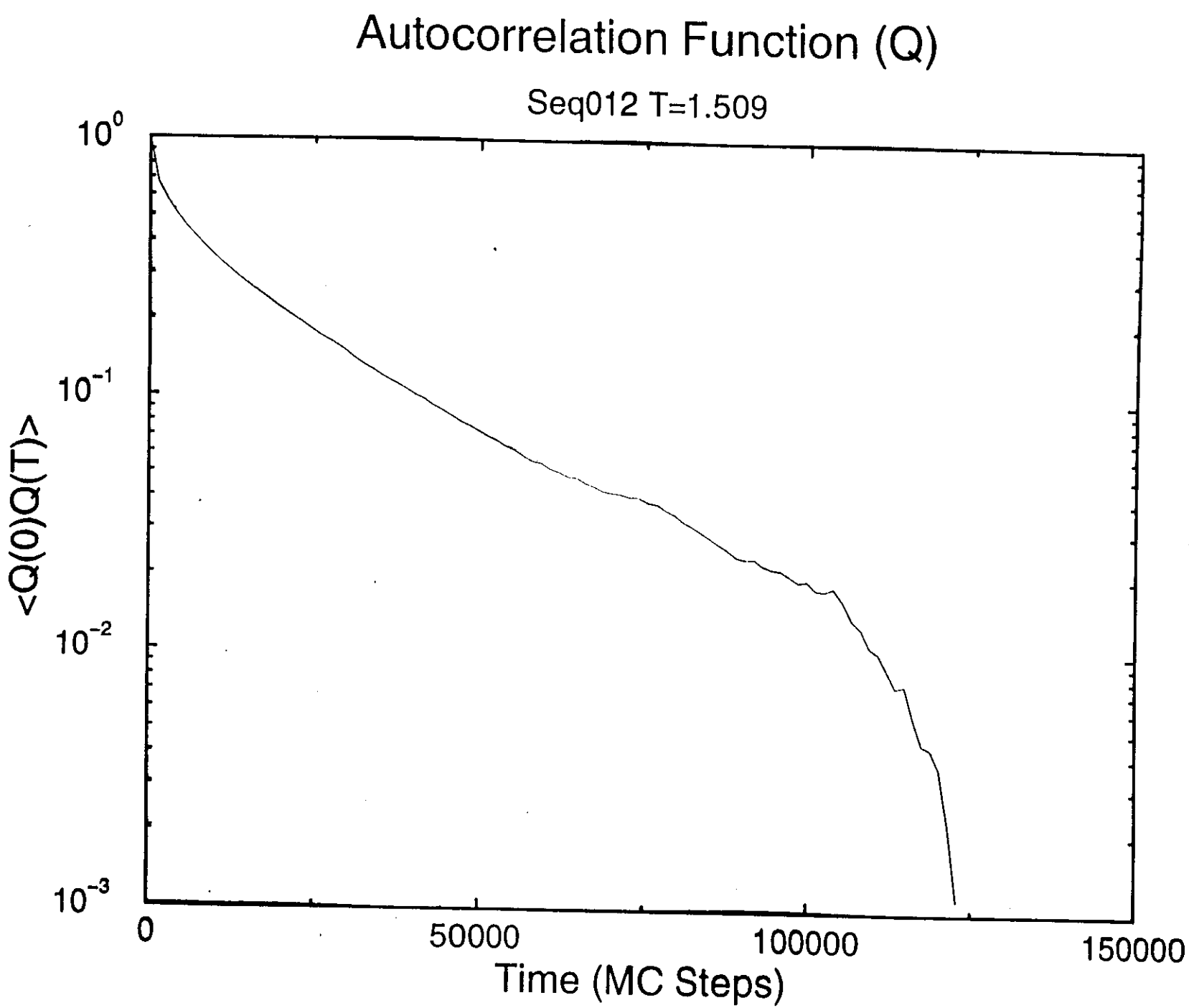
NDSocci

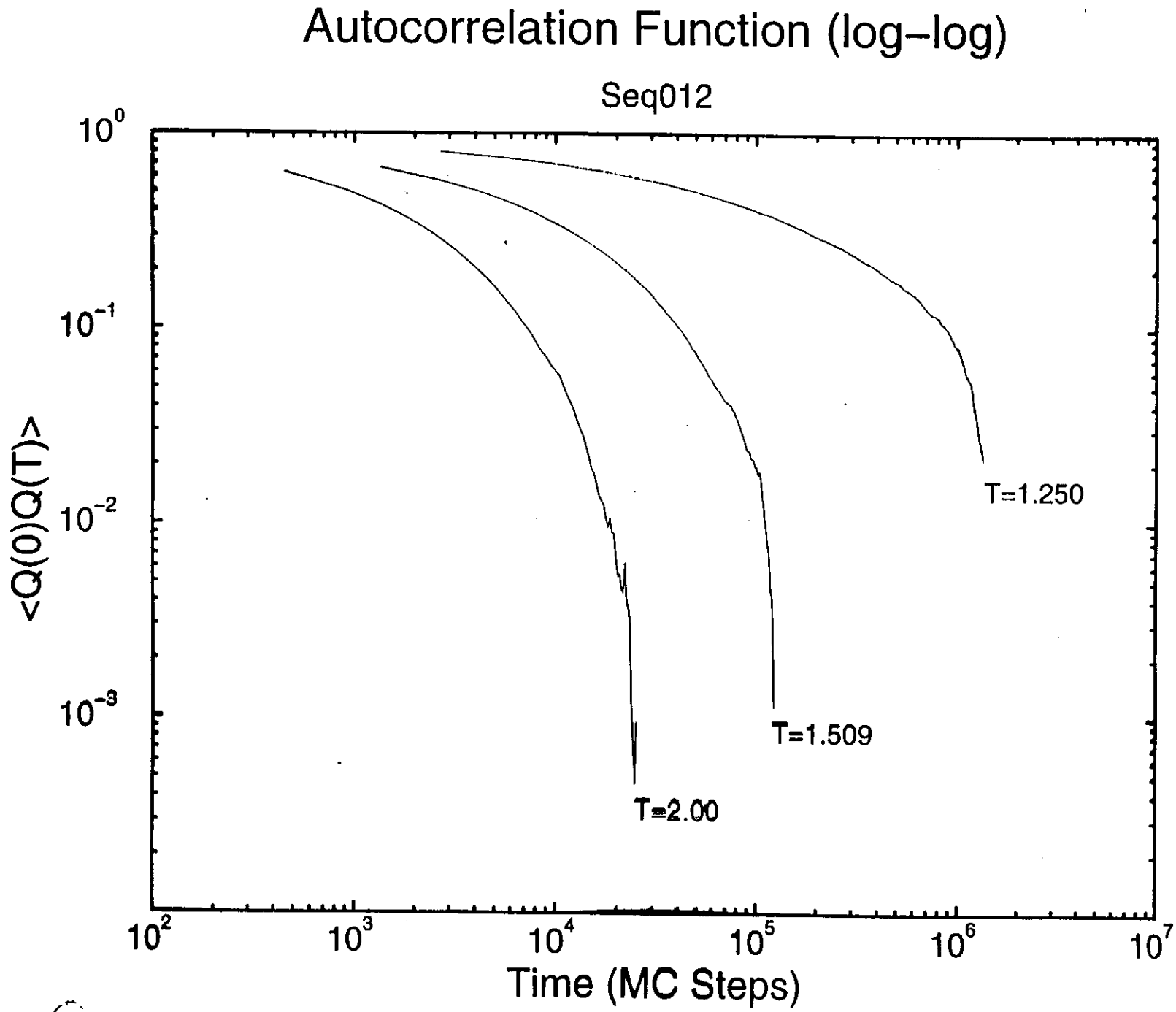




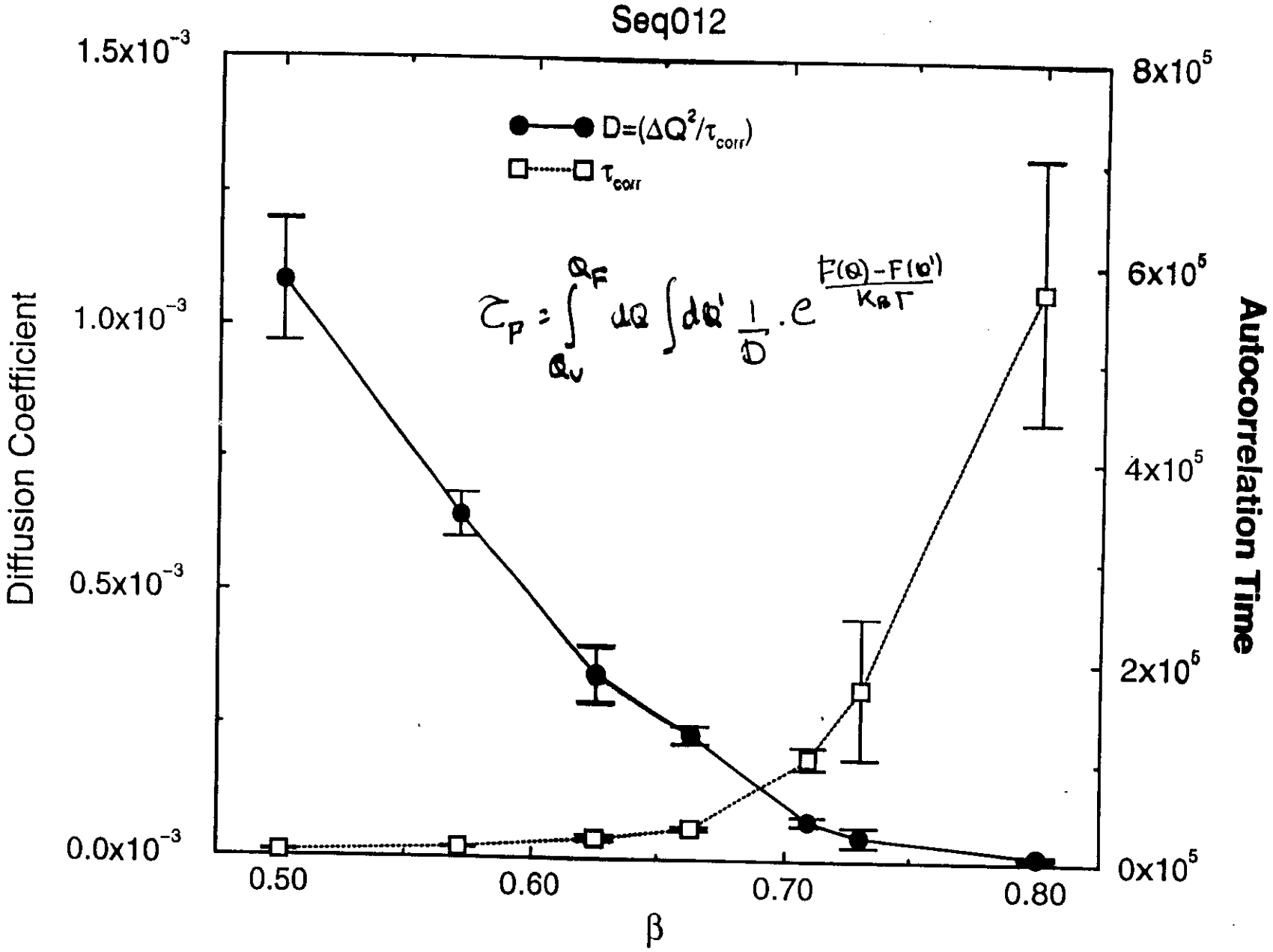
Very sharp

3000 steps


Autocorrelation Function (Q)
Seq012 T=1.509
10^0
10^{-1}
10^{-2}
10^{-3}
<Q(0)Q(T)>
0
50000
100000
150000
Time (MC Steps)

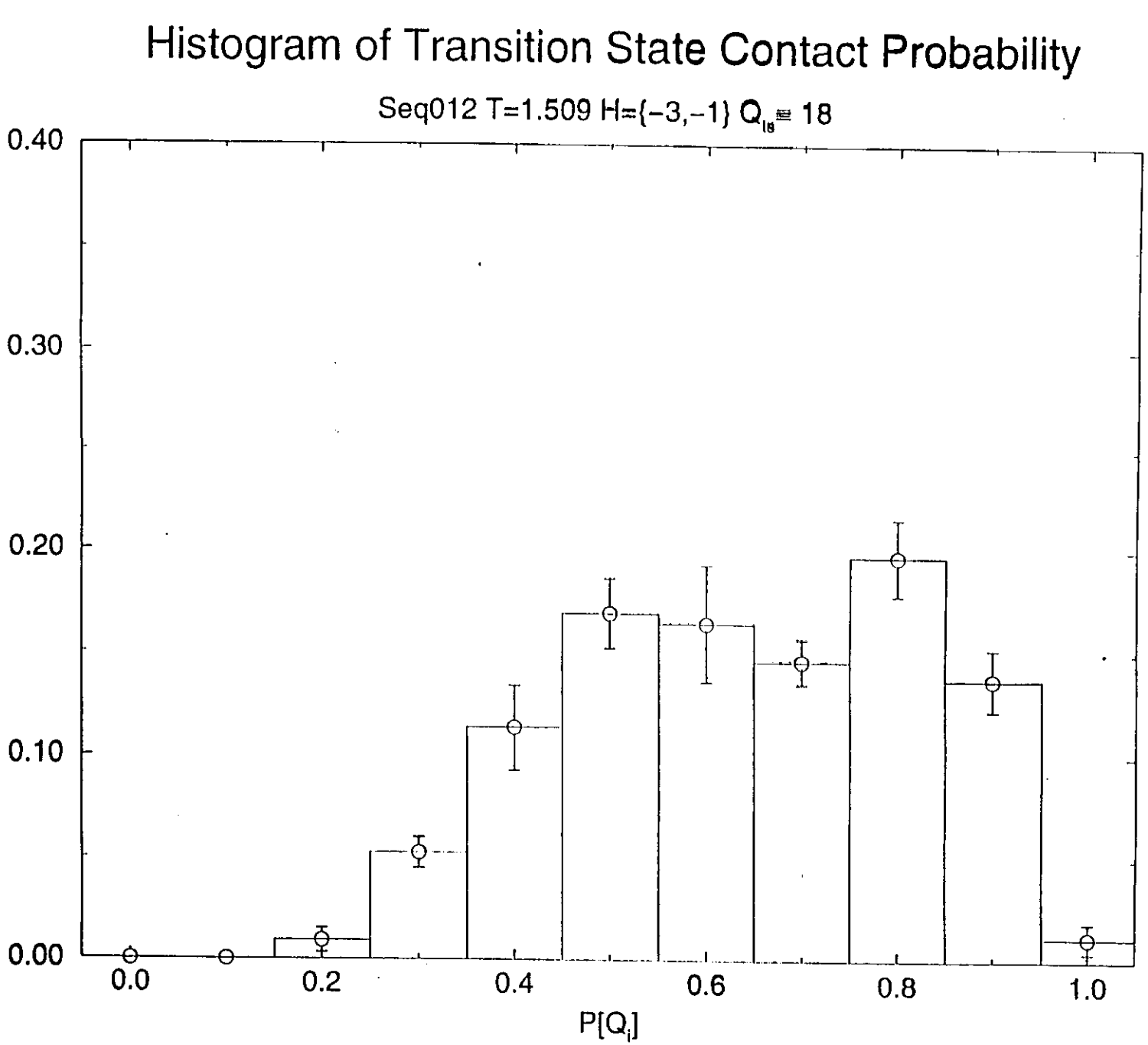

Autocorrelation Function (log–log)
Seq012
10^0
10^{-1}
10^{-2}
10^{-3}
<Q(0)Q(T)>
T=1.250
T=1.509
T=2.00
10^2
10^3
10^4
10^5
10^6
10^7
Time (MC Steps)

D and τ_{corr} vs Inverse Temperature

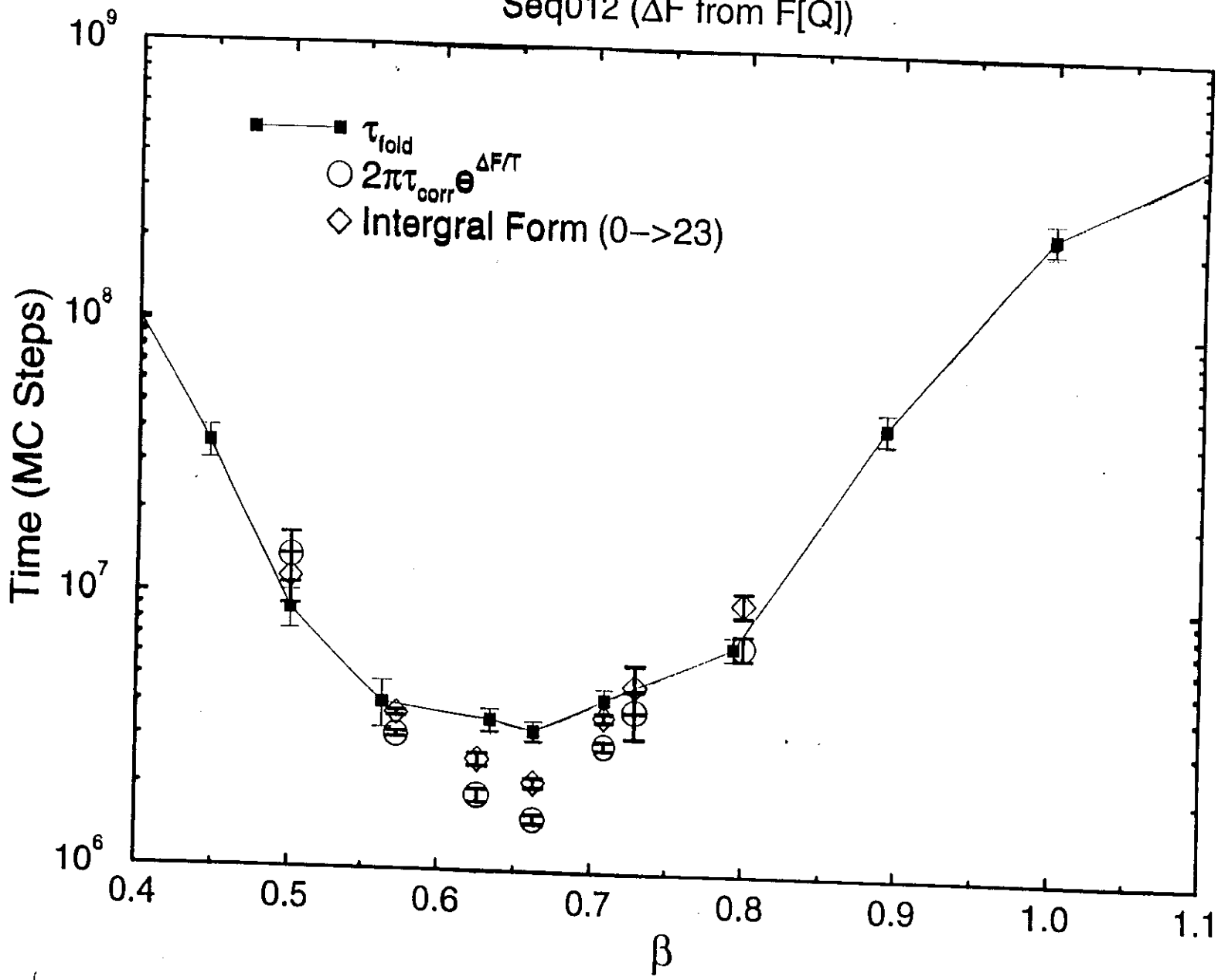


Mon Oct 16 21:47:25 1995

NDSocci

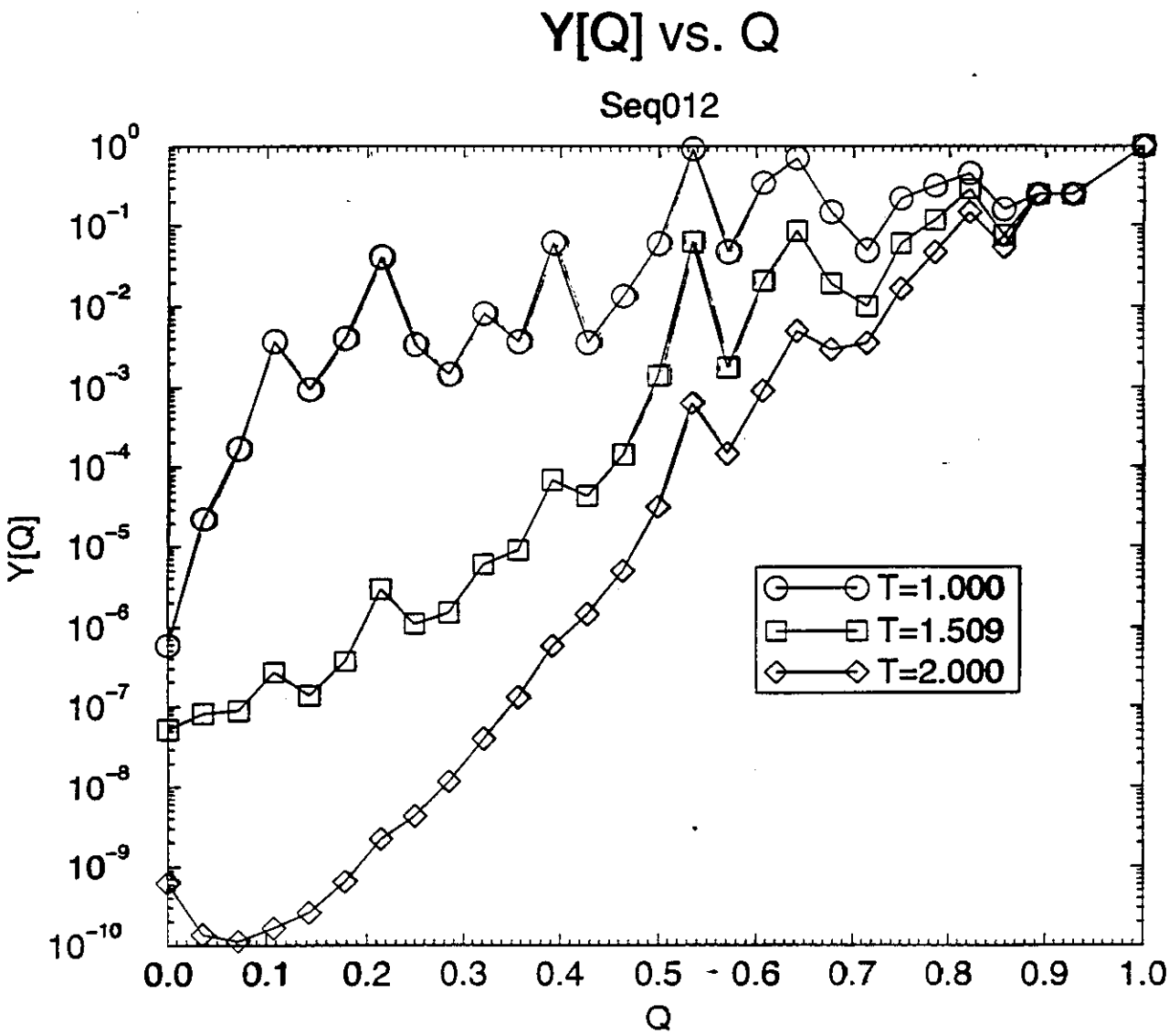

Histogram of Transition State Contact Probability
Seq012 T=1.509 H=(-3,-1) Q_{18}= 18
0.40
0.30
0.20
0.10
0.00
0.0
0.2
0.4
0.6
0.8
1.0
P[Q_i]

Kramers Rate vs MC MFPT



Thu Aug 10 22:46:24 1995

NDSocci

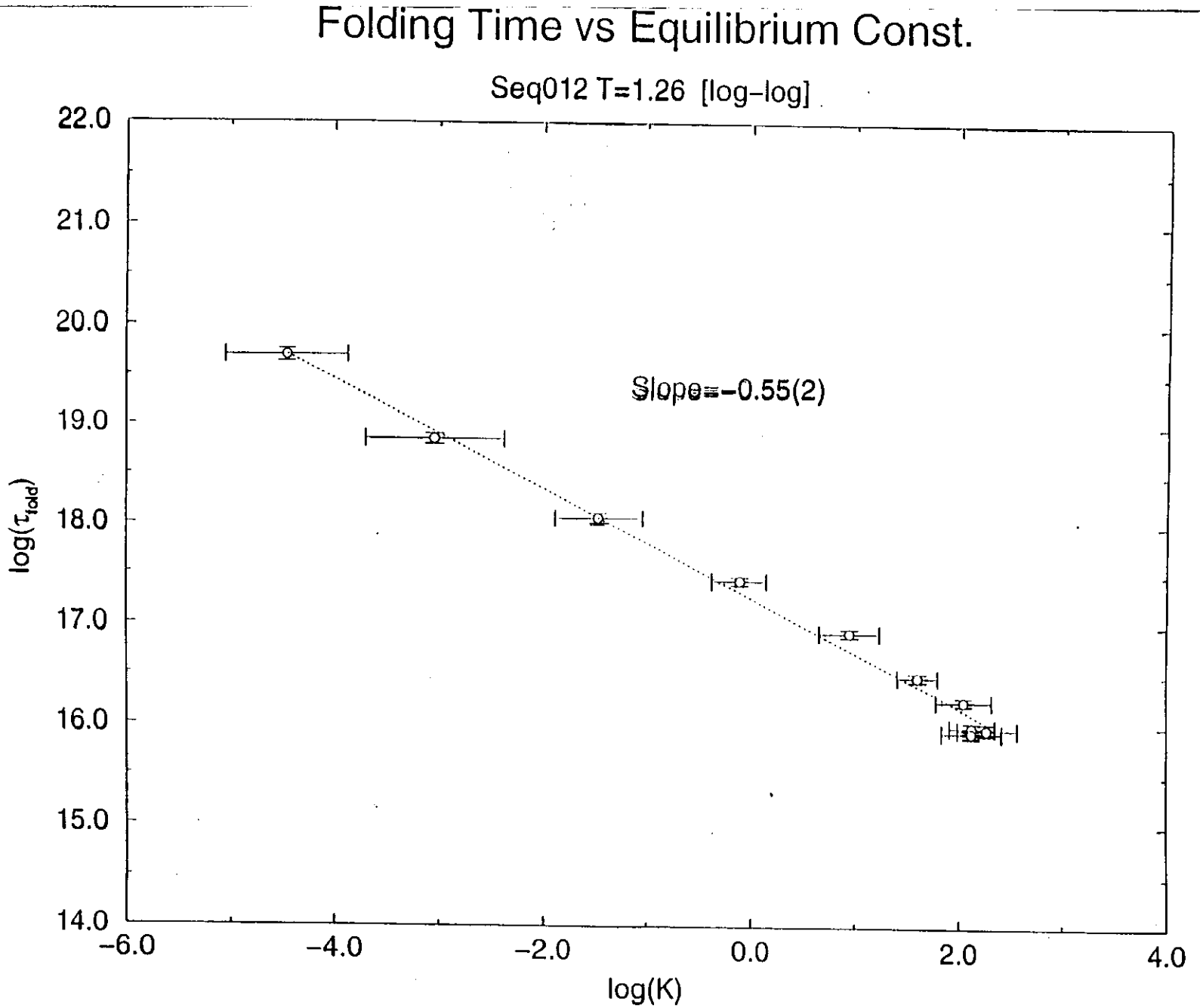

Y[Q] vs. Q
Seq012
Y[Q]
Q
T=1.000
T=1.509
T=2.000
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
10^-9
10^-10
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Folding Time

vs.

Denaturant Concentration

Theory and experiments



Sun Dec 17 23:48:13 1995

NDSocci

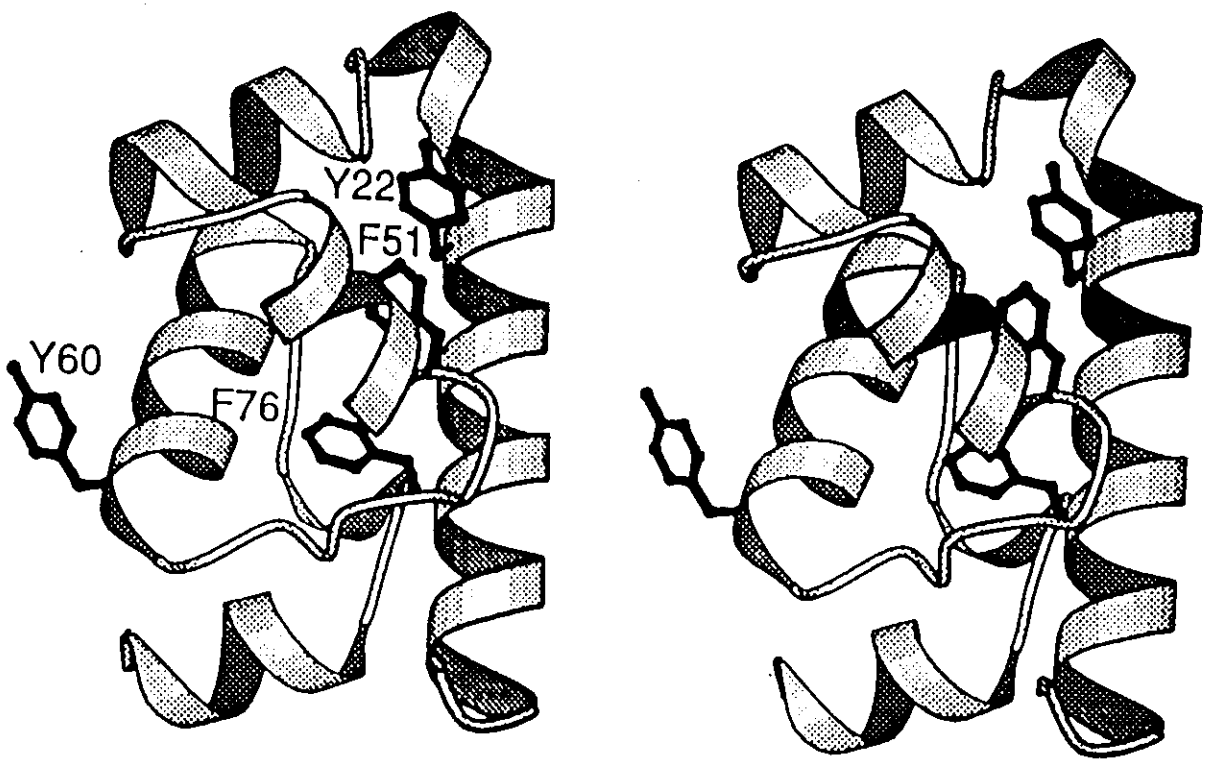


FIG. 1. Stereoview of λ_{6-85} coordinates from the protein–DNA cocrystal structure of Beamer and Pabo (27). Side chains of Tyr-22, Phe-51, Tyr-60, and Phe-76 are indicated. Figure was made with MOLSCRIPT (29).

urea concentrations, each molecule samples both states several times during the acquisition of the NMR spectrum. This exchange process leads to a shift in peak position and line broadening. As shown in Fig. 3, the aromatic ^{1}H resonances of

The spectrum of a protein resonance in two-state exchange between native (N) and denatured (D) states is given by the following equation for the intensity as a function of frequency (13):

D),

unction of
ortional to
requencies
k_u are the
$_D$ are the
e apparent
and D. The
C_0, $\delta\nu$, T_{2N},
ied directly
ıtzian line-
refore, the
simulating
mple of this
ıted in Fig.
sufficiently
parameters

Tyr-22 εH

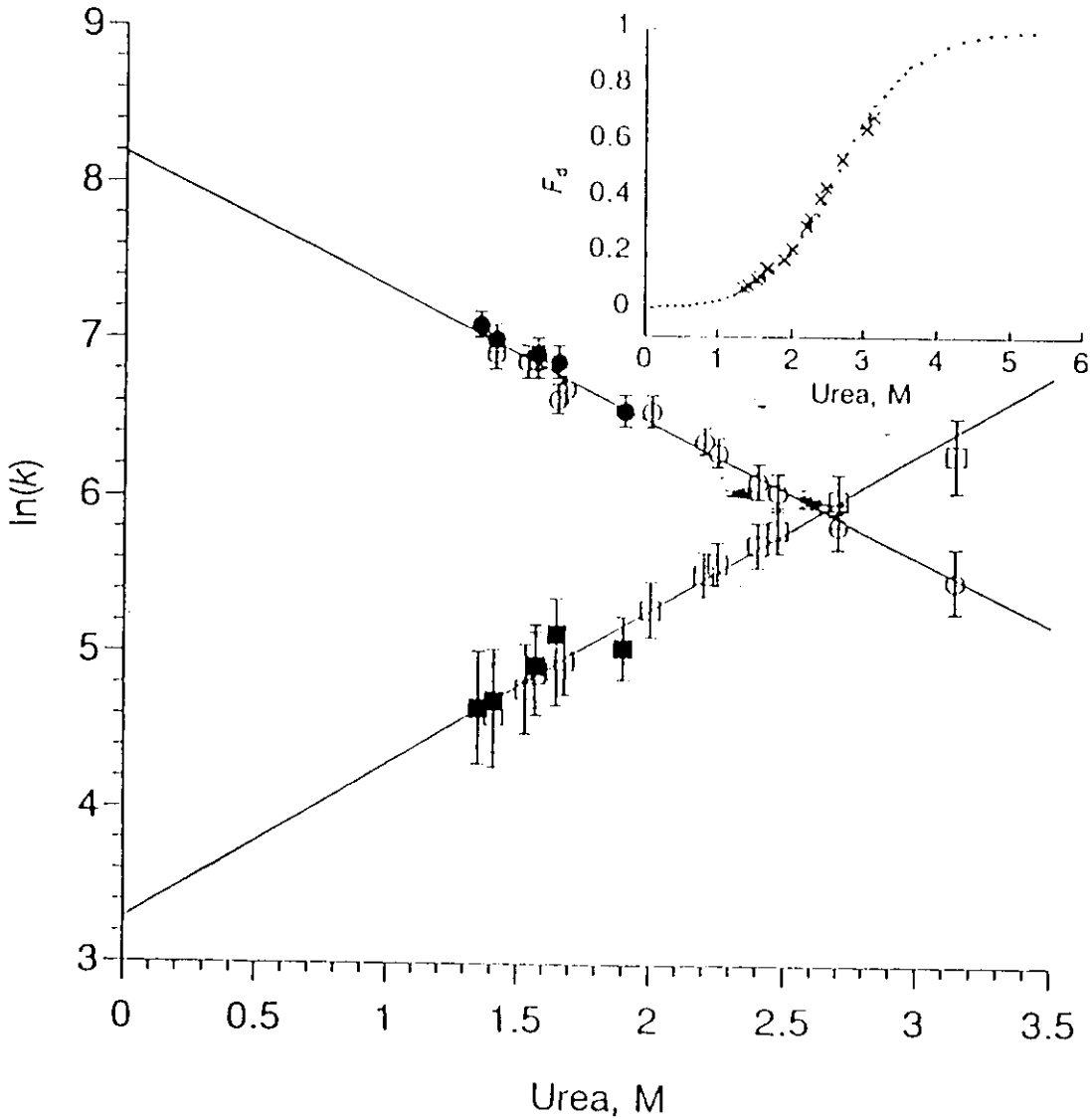


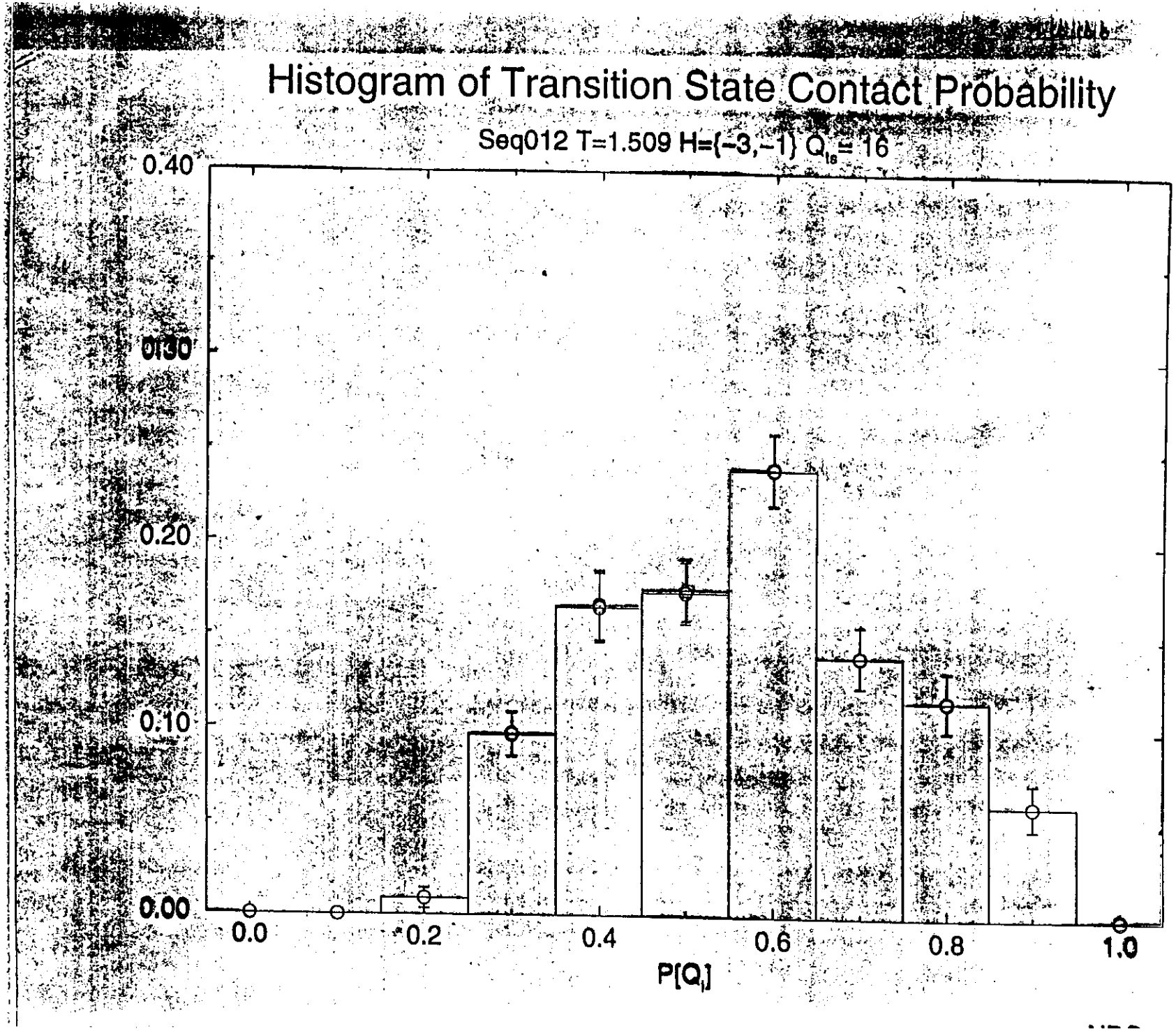
FIG. 5. Natural logarithm of the λ_{6-85} folding (circles) and unfolding (squares) rate constants vs. urea concentration at 37°C. Open

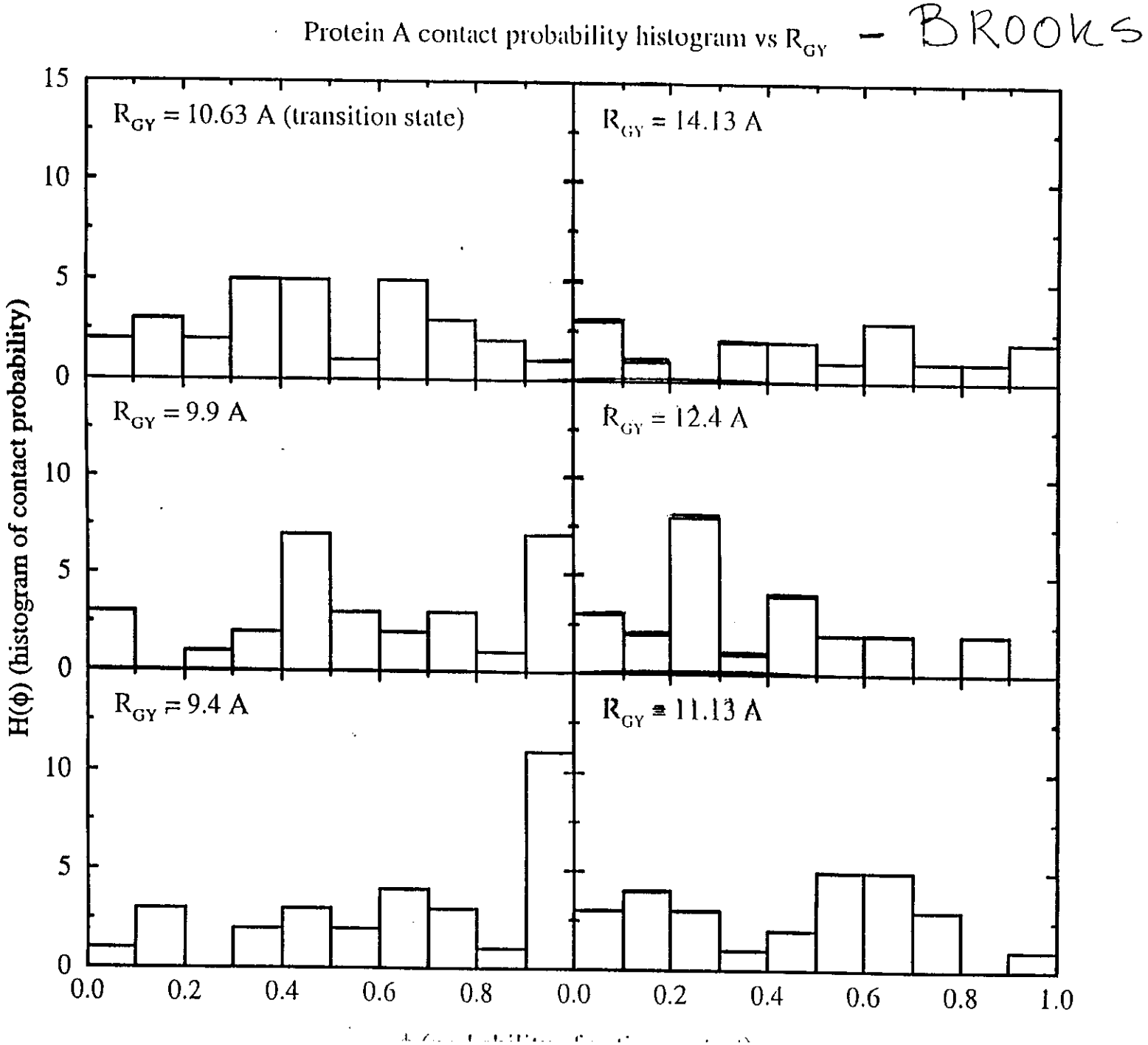
$\Delta G_f = \Delta G_0 + m[\text{Urea}]$

Slope ~ 0.48

log K

The Nature of the Transition State Ensemble


Histogram of Transition State Contact Probability
Seq012 T=1.509 H=(-3,-1) Q_ts = 16
0.40
0.30
0.20
0.10
0.00
0.0
0.2
0.4
0.6
0.8
1.0
P[Q_i]


Protein A contact probability histogram vs R_{GY} – BROOKS
R_{GY} = 10.63 A (transition state)
R_{GY} = 14.13 A
R_{GY} = 9.9 A
R_{GY} = 12.4 A
R_{GY} = 9.4 A
R_{GY} = 11.13 A
H(ϕ) (histogram of contact probability)
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0

393

Protein folding
funnel

A
B
C
D