Bioinformatique M2: Lecture 4 – part A

P. Derreumaux

III. From protein sequence to 3D structure

The Protein Folding problem

2 aspects

- structures from sequences.

- sequence events from disordered to native structures (characterization of TSE) →

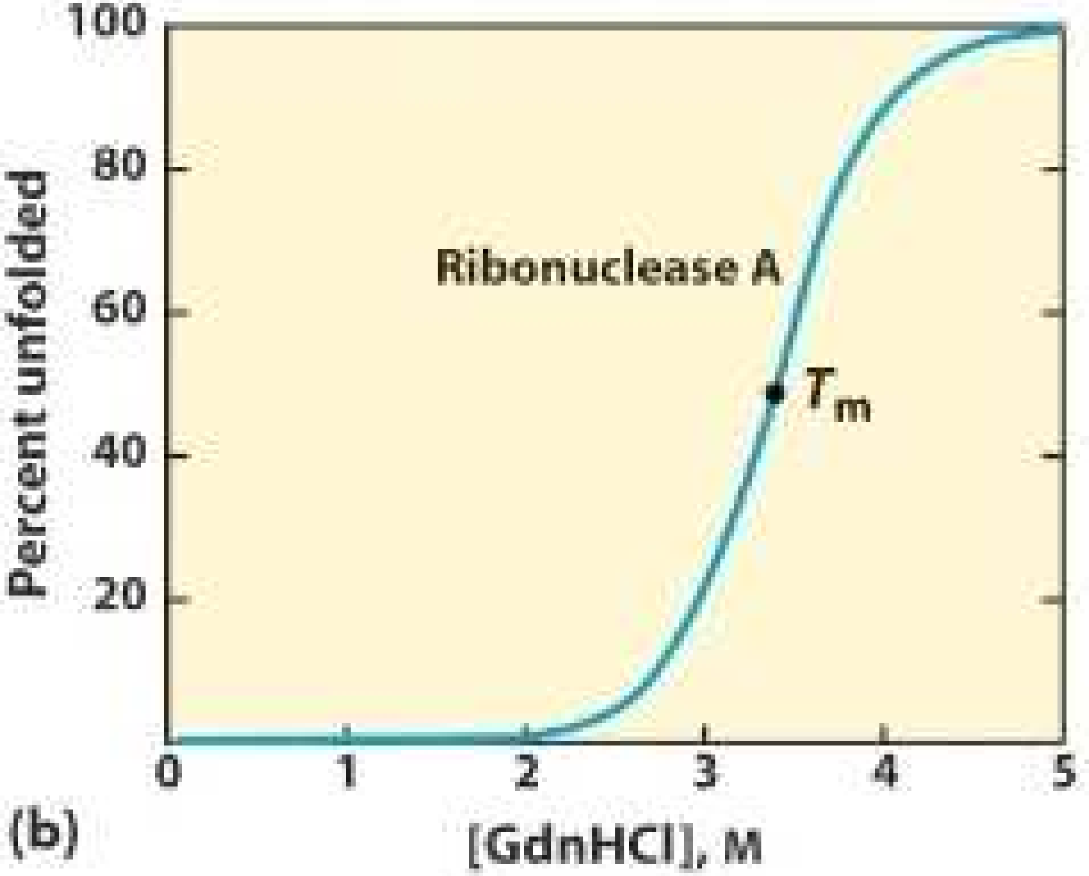
  sequence elements coding for protein topology

-Folding under chaperon control

-Thermodynamic hypothesis of folding (Anfinsen)

- Amyloid Fibril Formation

Monomeric Protein Folding: a simple experimental view



Monomeric Protein Folding: a simple view

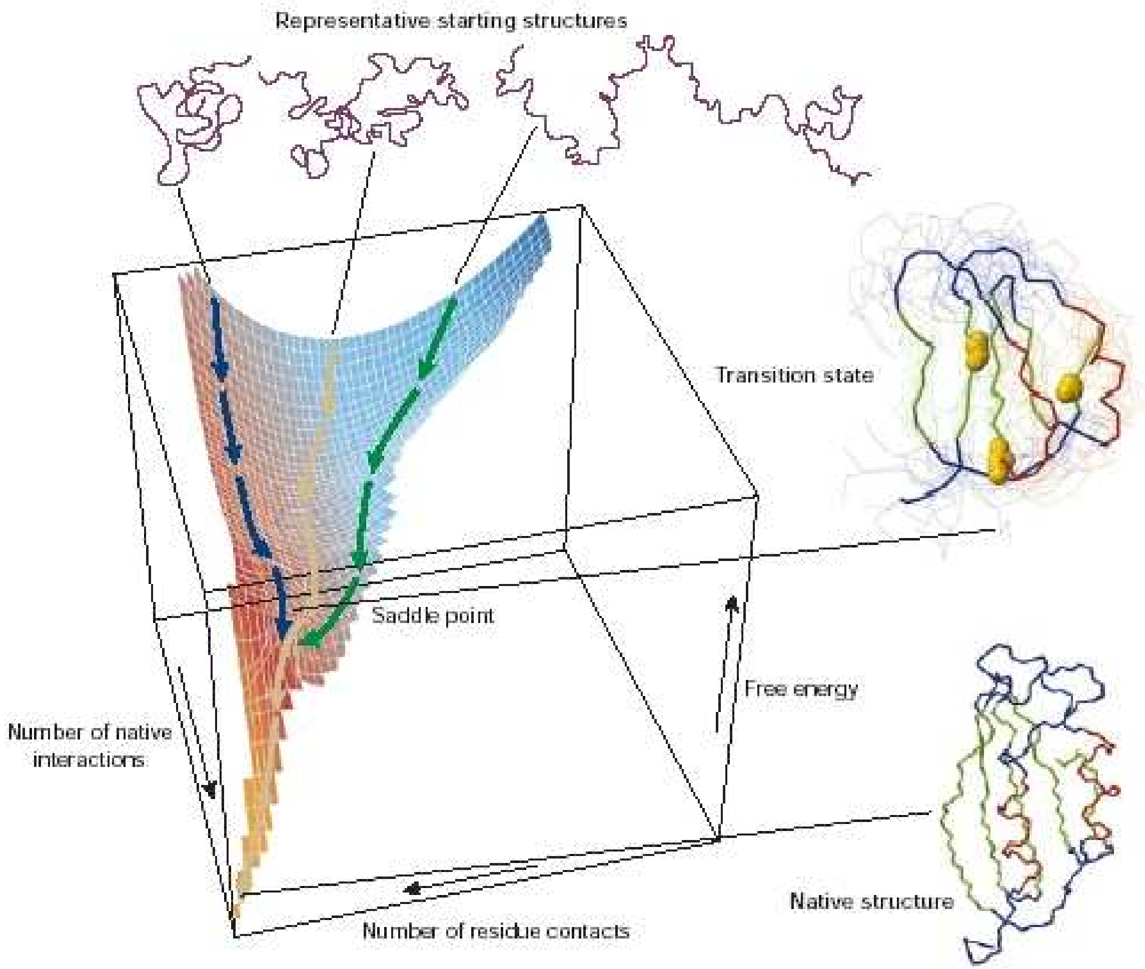
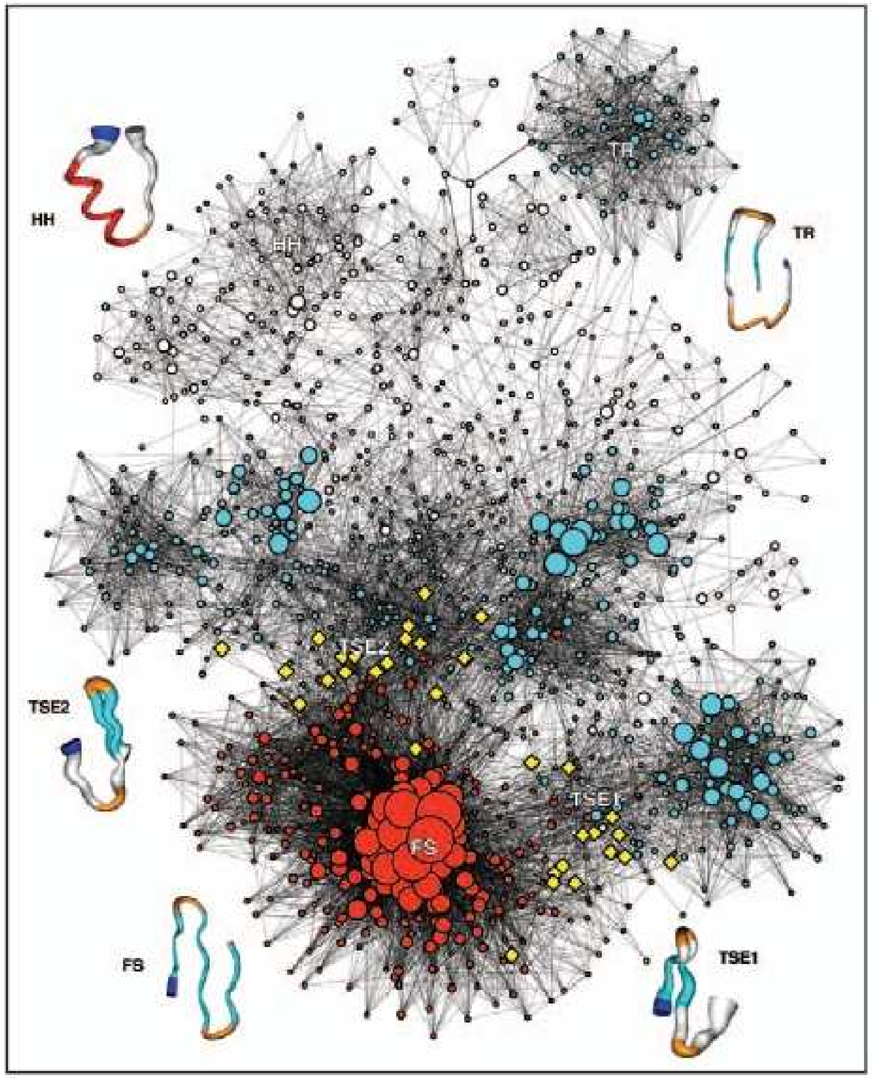

Representative starting structures
Transition state
Saddle point
Free energy
Number of native interactions
Number of residue contacts
Native structure

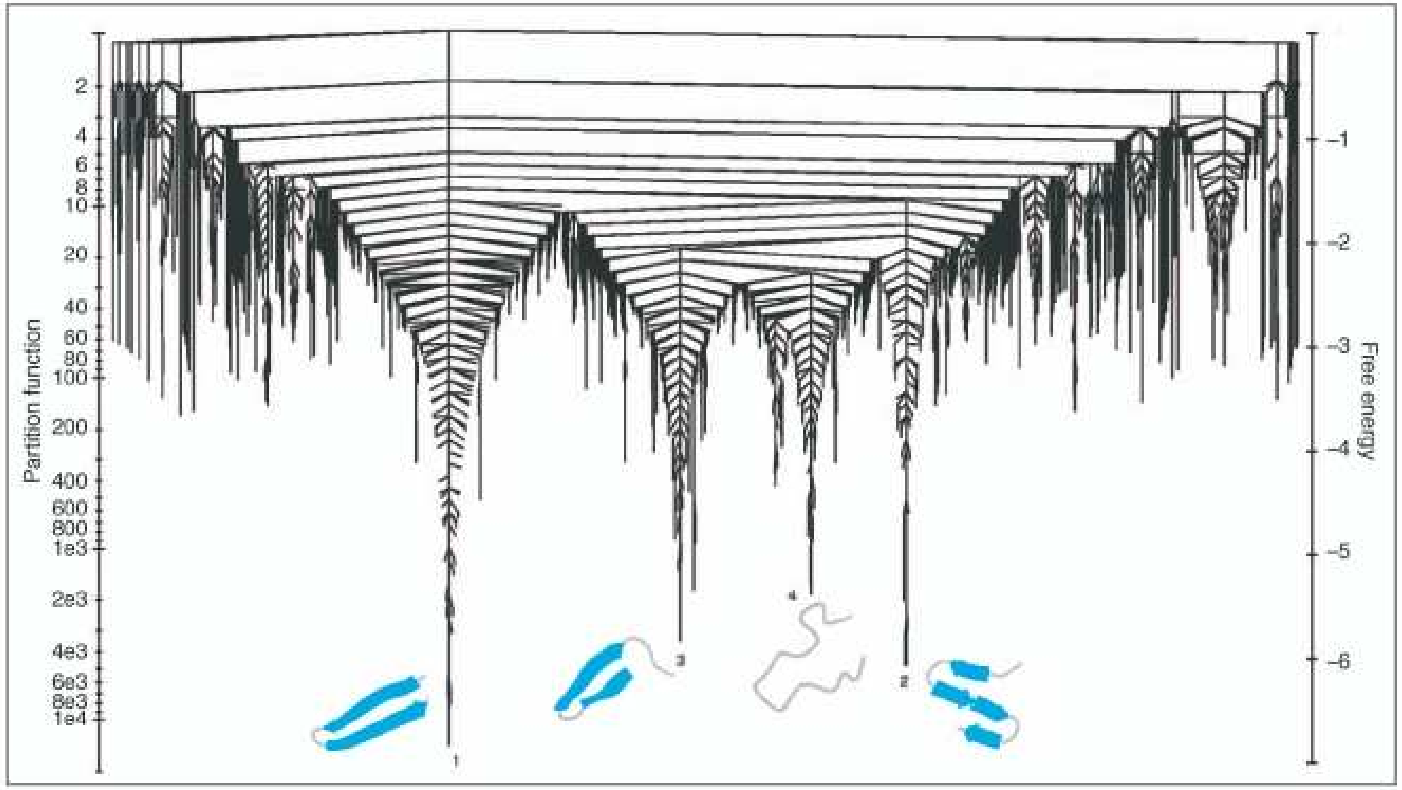
Figure 1 A schematic energy landscape for protein folding. The surface is derived from a computer simulation of the folding of a highly simplified model of a small protein. The surface 'funnels' the multitude of denatured conformations to the unique native structure. The critical region on a simple surface such as this one is the saddle point corresponding to the transition state, the barrier that all molecules must cross if they are to fold to the native state. Superimposed on this schematic surface are ensembles of structures corresponding to different stages of the folding process. The transition state ensemble was calculated by using computer simulations constrained by experimental data from mutational studies of acylphosphatase[18]. The yellow spheres in this ensemble represent the three 'key residues' in the structure; when these residues have formed their native-like contacts the overall topology of the native fold is established. The structure of the native state is shown at the bottom of the surface; at the top are indicated schematically some contributors to the distribution of unfolded species that represent the starting point for folding. Also indicated on the surface are highly simplified trajectories for the folding of individual molecules. Adapted from ref. 6.

The funnel energy surface is oversimplified

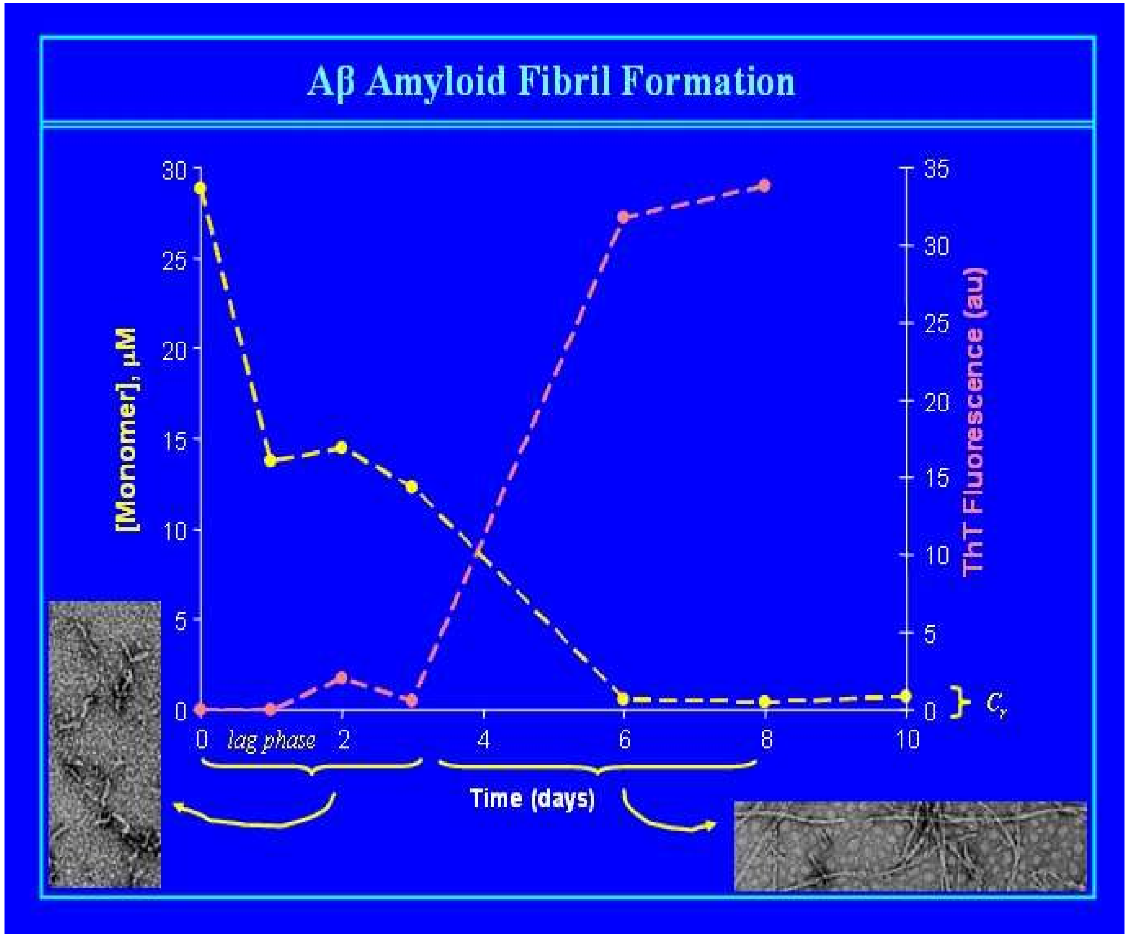


The funnel energy surface is oversimplified

Figure 3



Transition disconnectivity graph of a β-hairpin (the C-terminal segment from the B1 domain of protein G). A total of 4 μs implicit solvent molecular dynamics simulations at 360 K were sampled to obtain a sufficient number of folding-unfolding events [29]. Representative structures of the deepest free energy minima are shown and labeled 1–4. The left vertical axis shows the partition function of the minima and barriers. The right vertical axis shows the free energy of the minima and barriers. Reproduced with permission from [29].



Aβ Amyloid Fibril Formation

amyloid fibril formation is described by a polymerisation-nucleation process

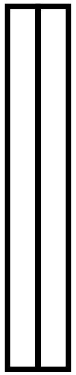
Details of the polymerization-nucleation process

LAG PHASE

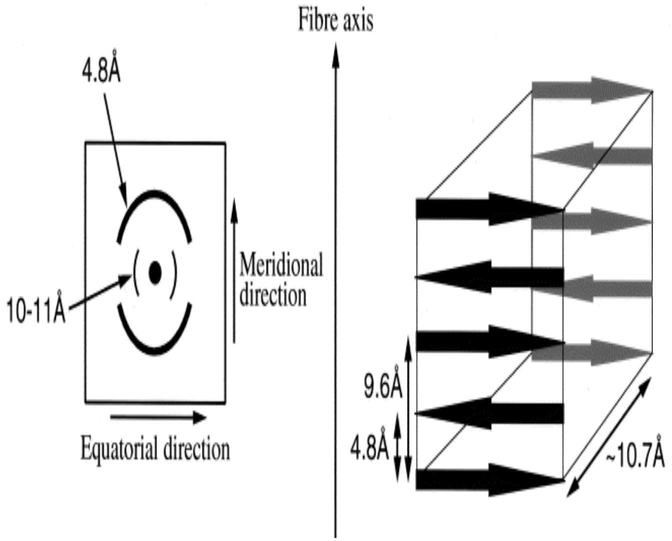
NUCLEUS

Monomers ⇌ 2-,3-mers ⇌ Spherical, Annular, Amorphous Forms → ?? (*Aβ40/42*) 12-20 mers (PrP)

GROWTH



fibrils



Overall, aggregation very sensitive to sequence, pH, concentration, anions

Prediction 3D structure from sequence

-Genomic Programs and the number of sequences

-Structures are more conserved than sequences

-Experimental costs for structure determination

The Protein Folding problem and A limited number of folds. ($\sim 800 < N_{fold} < 5000$)

<u>basic reasons</u>:

<u>obvious</u> — structural stability

$$E_{Nat} \rightarrow P(E_{Nat}) = \frac{e^{-\frac{E_{Nat}}{KT}}}{\sum_i e^{-\frac{E_i}{KT}}} \gg P(E_i$$

— functional property
(N.B. Identical function for different folds

<u>not obvious</u> — kinetic property
('fast folding seq')
'folding nucleus'

— AA mutation variability and its correlation with the number of folds

— AA mutation and its effect on structural stability.

Coût d’une structure 3D

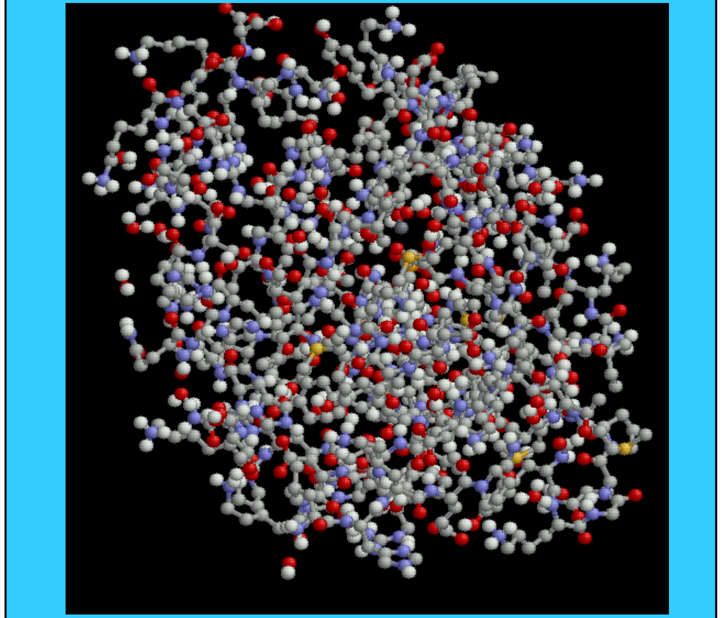
	Coût moyen	Change de succès
Protéines solubles bactériennes	$140000	35%
Protéines humaines solubles (kinases, protéases, ...)	$450000	35%
Protéines membranaires bactériennes	$1,5 million	10%
Protéines membranaires humaines	$2,5 million	10%

R.C. Stevens, Drug Discovery, 2003

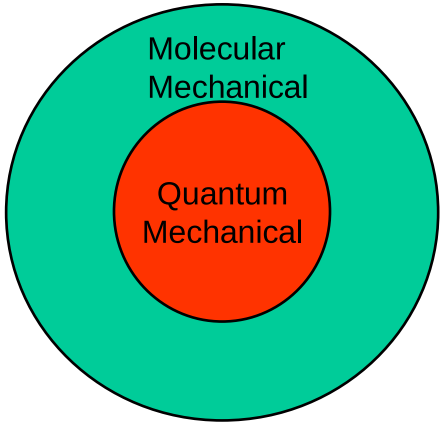
Coût n’incluant pas les développements technologiques ni l’amortissement des équipements lourds

Computer Simulation - Basic Principles

Model System

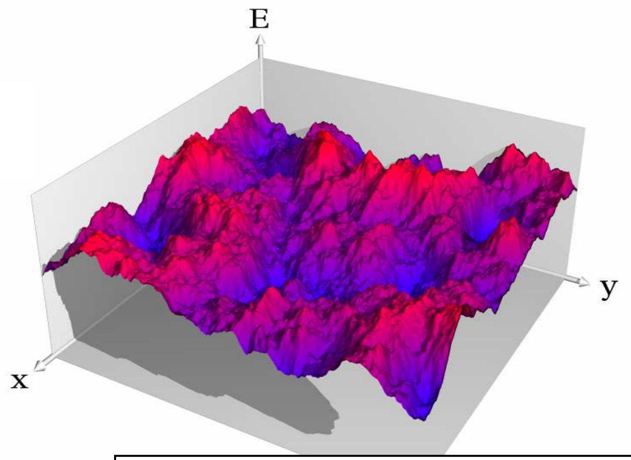


Coarse-grained Force field



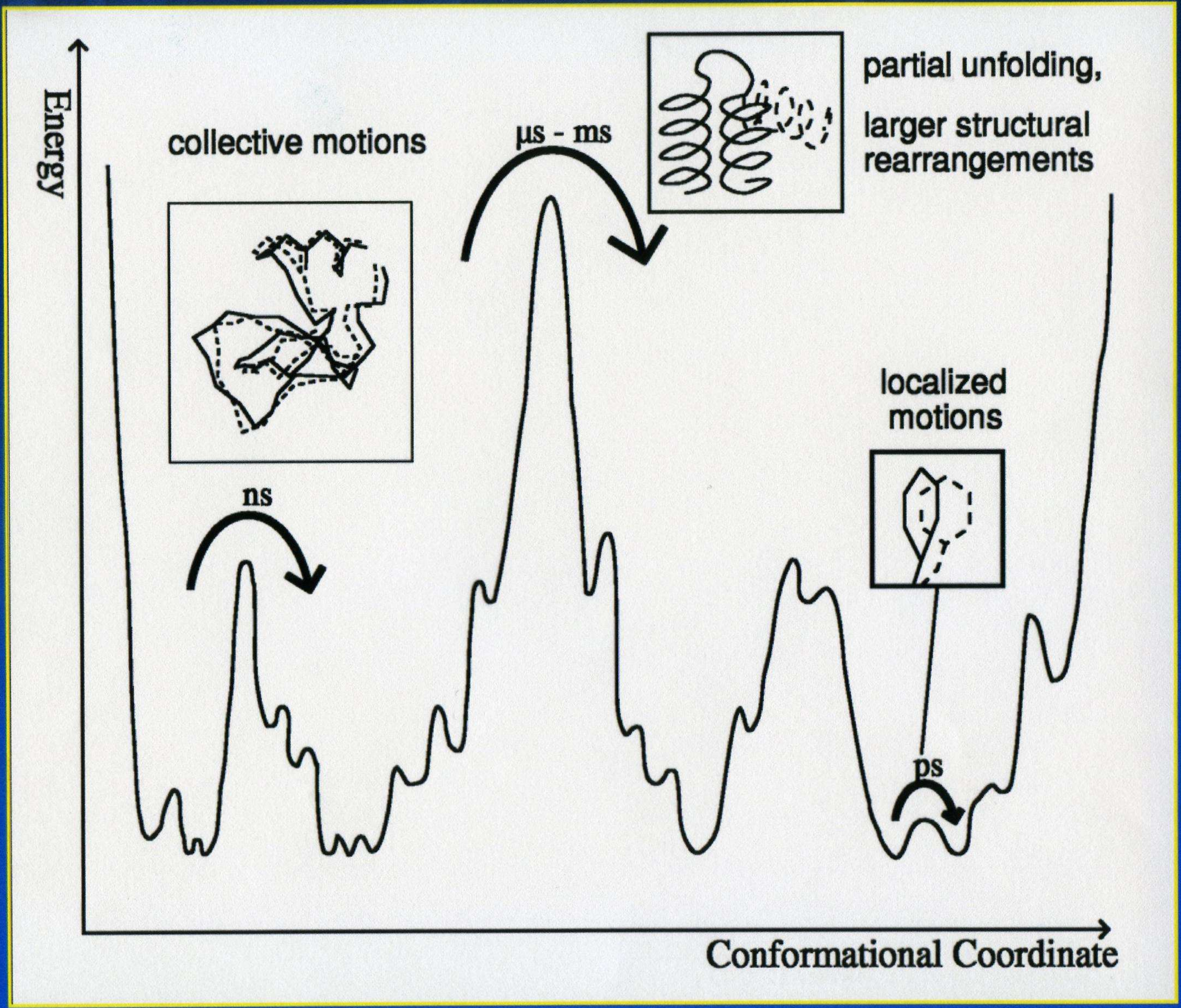
Molecular Mechanical

Quantum Mechanical



Techniques to explore the energy landscape (structures, thermodynamics)

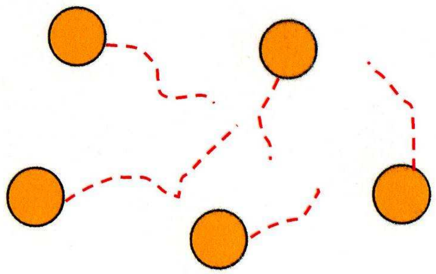
Proteins jump between many, hierarchically ordered ***conformational substates***



H. Frauenfelder et al., *Science* **229** (1985) 337

What is a molecular dynamics simulation?

- Simulation that shows how the atoms in the system move with time
- Typically on the nanosecond timescale
- Atoms are treated like hard balls, and their motions are described by Newton's laws.



Ideally : MD simulations in solvent.

MD simulation : a simple application.

$$F = m a = m \frac{dv}{dt} = m \frac{d^2x}{dt^2}$$

Supp $\quad a = cte = \dfrac{dv}{dt}$

$$v = at + v_0$$

$$x = vt + x_0 = a t^2 + v_0 t + x_0$$

with $a = -\dfrac{1}{m} \dfrac{dV}{dx}$.

To calculate a trajectory, one needs.

- initial positions of the particles
- initial distribution of velocities
- the gradient of the potential energy function

Molecular Dynamics Simulation

Molecule: (classical) N-particle system

Newtonian equations of motion:

$$m_i \frac{d^2}{dt^2}\vec{r}_i = \vec{F}_i(\bar{r})$$

$$\vec{F}_i(\bar{r}) = -\nabla_i V(\bar{r})$$

with $\bar{r} = (\vec{r}_1,\ldots,\vec{r}_N)$

Integrate numerically via the „leapfrog“ scheme

$$\begin{aligned} \boldsymbol{v}\left(t+\frac{\Delta t}{2}\right) &= \boldsymbol{v}\left(t-\frac{\Delta t}{2}\right) + \frac{\boldsymbol{F}(t)}{m}\Delta t \\ \boldsymbol{r}(t+\Delta t) &= \boldsymbol{r}(t) + \boldsymbol{v}\left(t+\frac{\Delta t}{2}\right)\Delta t \end{aligned}$$

with

$\Delta t \approx 1$fs!

(equivalent to the Verlet algorithm)

How do you run a MD simulation?

- **Get the initial configuration**

  From x-ray crystallography or NMR spectroscopy (PDB)

- **Assign initial velocities**

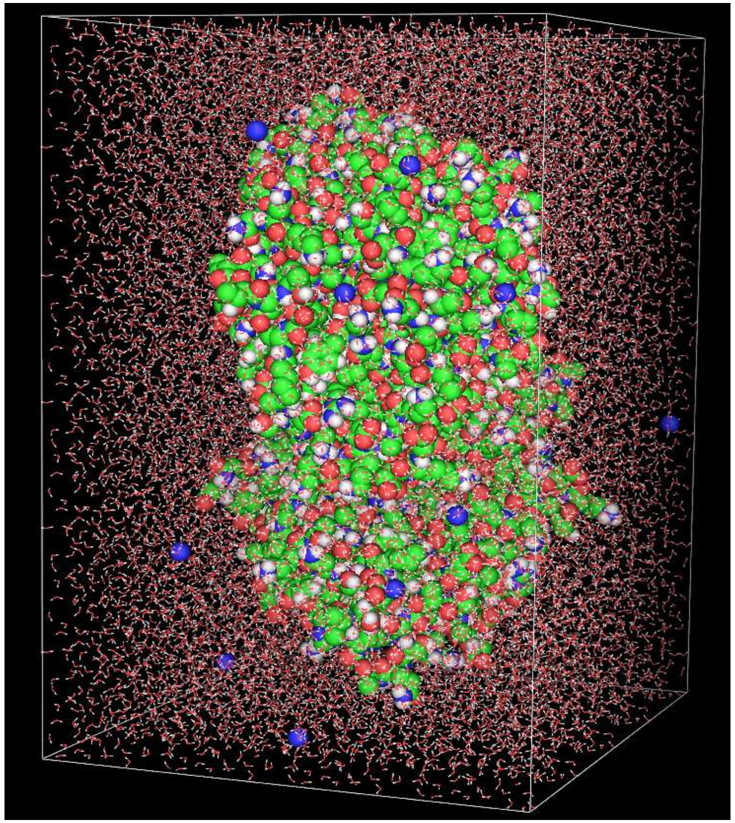
  At thermal equilibrium, the expected value of the kinetic energy of the system at temperature T is:

$$\left\langle E_{kin} \right\rangle = \frac{1}{2} \sum_{i=1}^{3N} m_i v_i^2 = \frac{1}{2} (3N) k_B T$$

  This can be obtained by assigning the velocity components vi from a random Gaussian distribution
  with mean 0 and standard deviation ($k_B T/m_i$):

$$\left\langle v_i^2 \right\rangle = \frac{k_B T}{m_i}$$

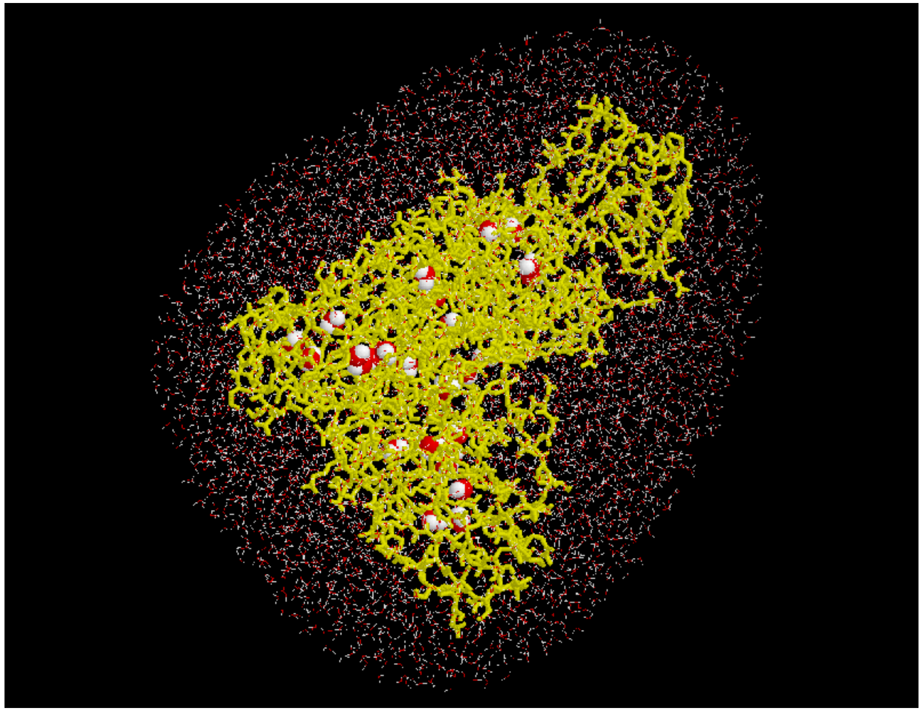
Treatment solvent



explicit

or

implicit?



Box or

Droplet ?

SPC, TIP3P..?

$\Delta t = 1\text{-}2\,fs$

Molecular dynamics

36 amino acids
+3,000 water
molecules

1-μs simulation
(exptl. 5-10 μs)

256 parallel processors
on a CRAY T3E

2 months of computer
time

Y.Duan and P.A.Kollman, Science, 282,740 (1998)

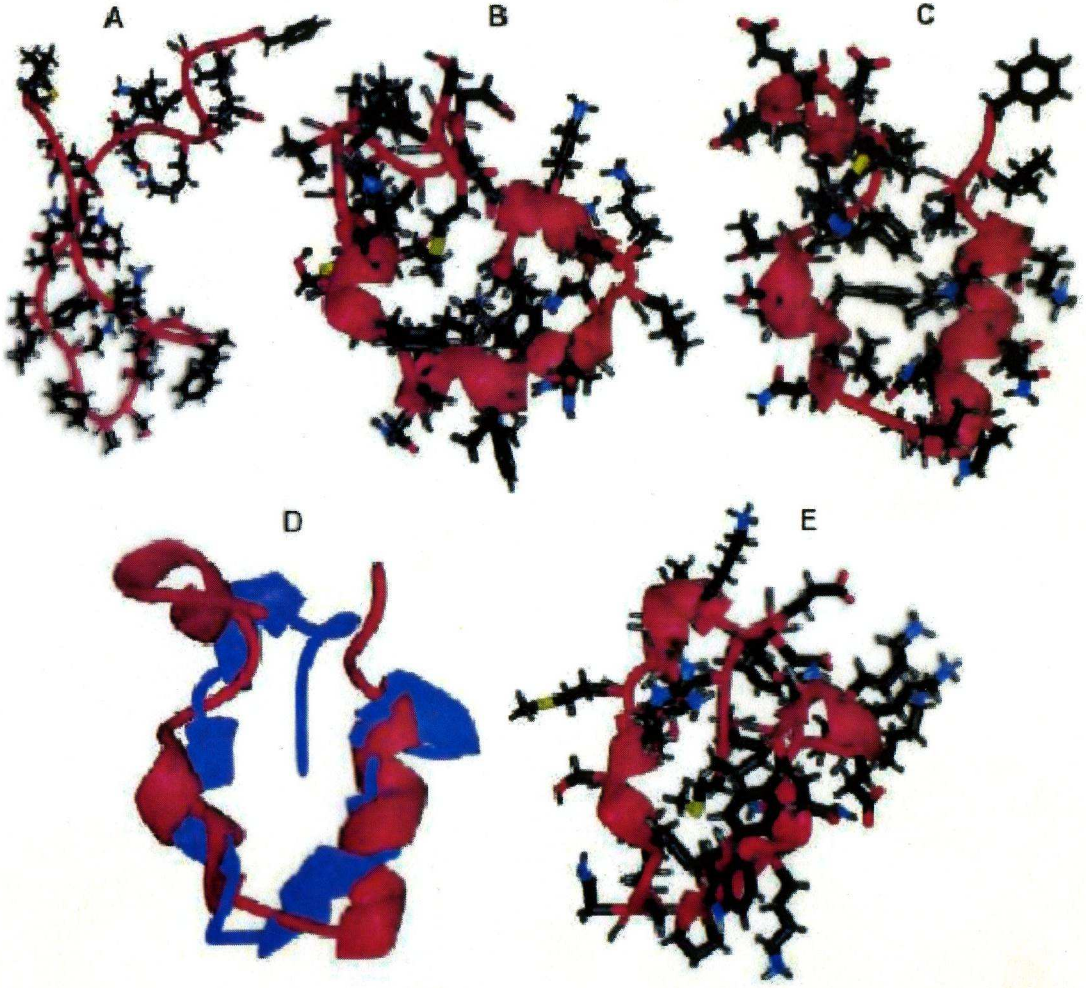
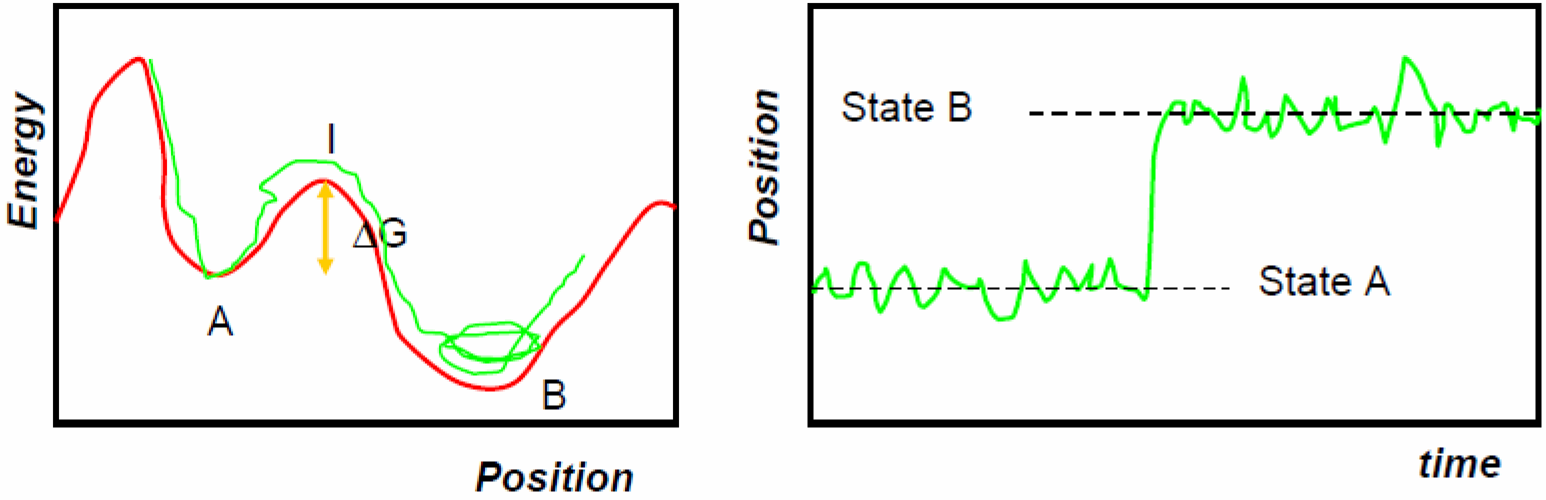


Fig. 1. Ribbon representations of (A) the unfolded, (B) partially folded (at 980 ns), and (C) native structures, and (E) a representative structure of the most stable cluster and (D) the overlap of the native (red) and the most stable cluster (blue) structures, generated with UCSF MidasPlus. Color code [except (D)]: red, main chain atoms and oxygen; black, non-main chain carbon; blue, non-main chain nitrogen; gray, hydrogen; yellow, sulfur.

Crossing energy barriers



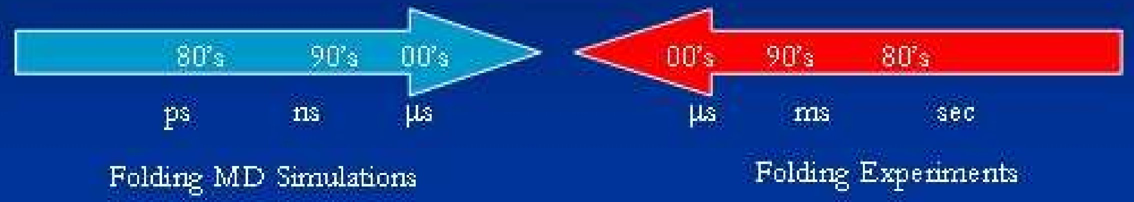
The actual transition time from A to B is very quick (a few pico seconds).

What takes time is waiting. The average waiting time for going from A to B can be expressed as:

$$\tau_{A\rightarrow B} = Ce^{\frac{\Delta G}{kT}}$$

Protein Folding: Fast Folders

Time Scale:



80's 90's 00's

ps ns µs

Folding MD Simulations

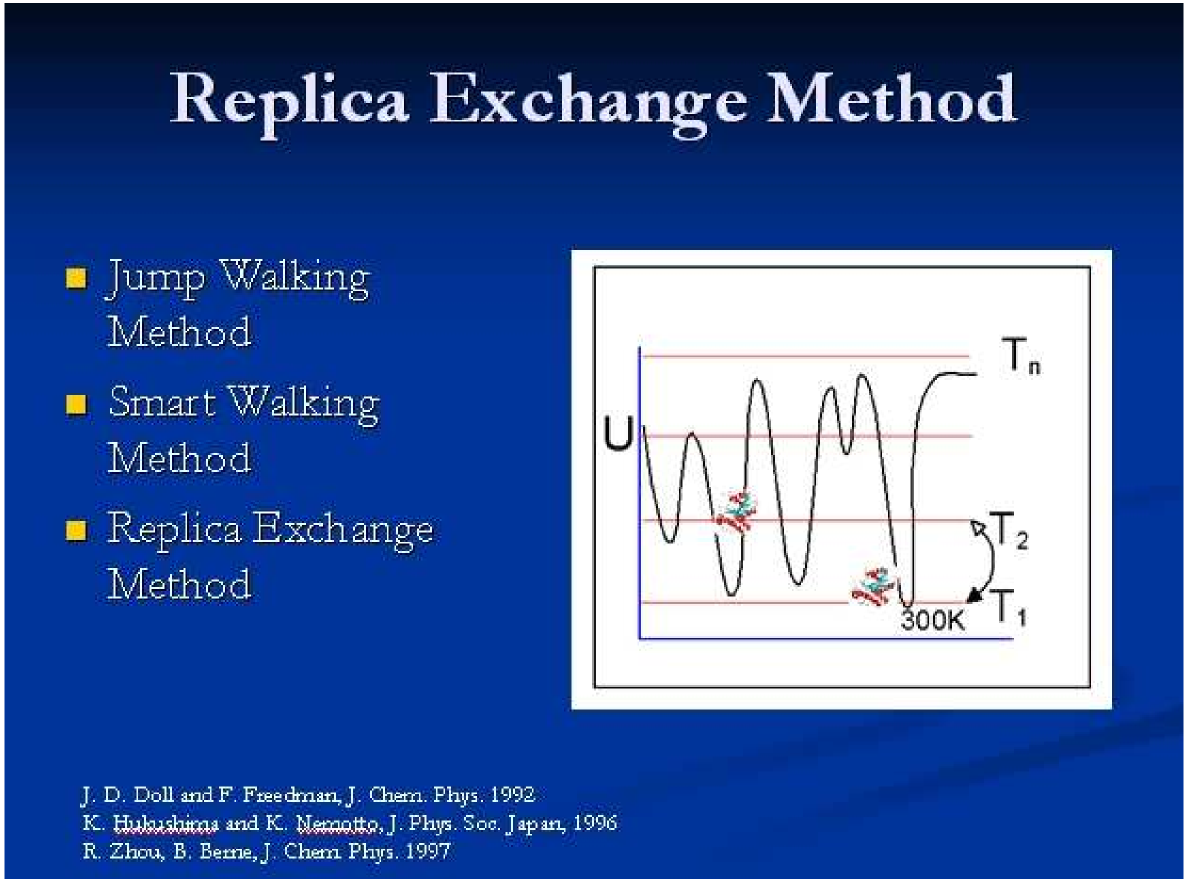
00's 90's 80's

µs ms sec

Folding Experiments

- Trp-cage, designed mini-protein (20 aa): 4µs
- β-hairpin of C-terminus of protein G (16 aa) : 6µs
- Engrailed homeodomain (En-HD) (61 aa): ~27µs
- WW domains (38-44 aa): >24µs
- Fe(II) cytochrome b_{562} (106 aa): extrapolated ~5µs
- B domain of protein A (58 aa): extrapolated ~8µs

One advanced technique



Replica Exchange Method

- Jump Walking Method
- Smart Walking Method
- Replica Exchange Method

U
T_n
T_2
T_1
300K

J. D. Doll and F. Freedman, J. Chem. Phys. 1992
K. Hukushima and K. Nemotto, J. Phys. Soc. Japan, 1996
R. Zhou, B. Berne, J. Chem. Phys. 1997