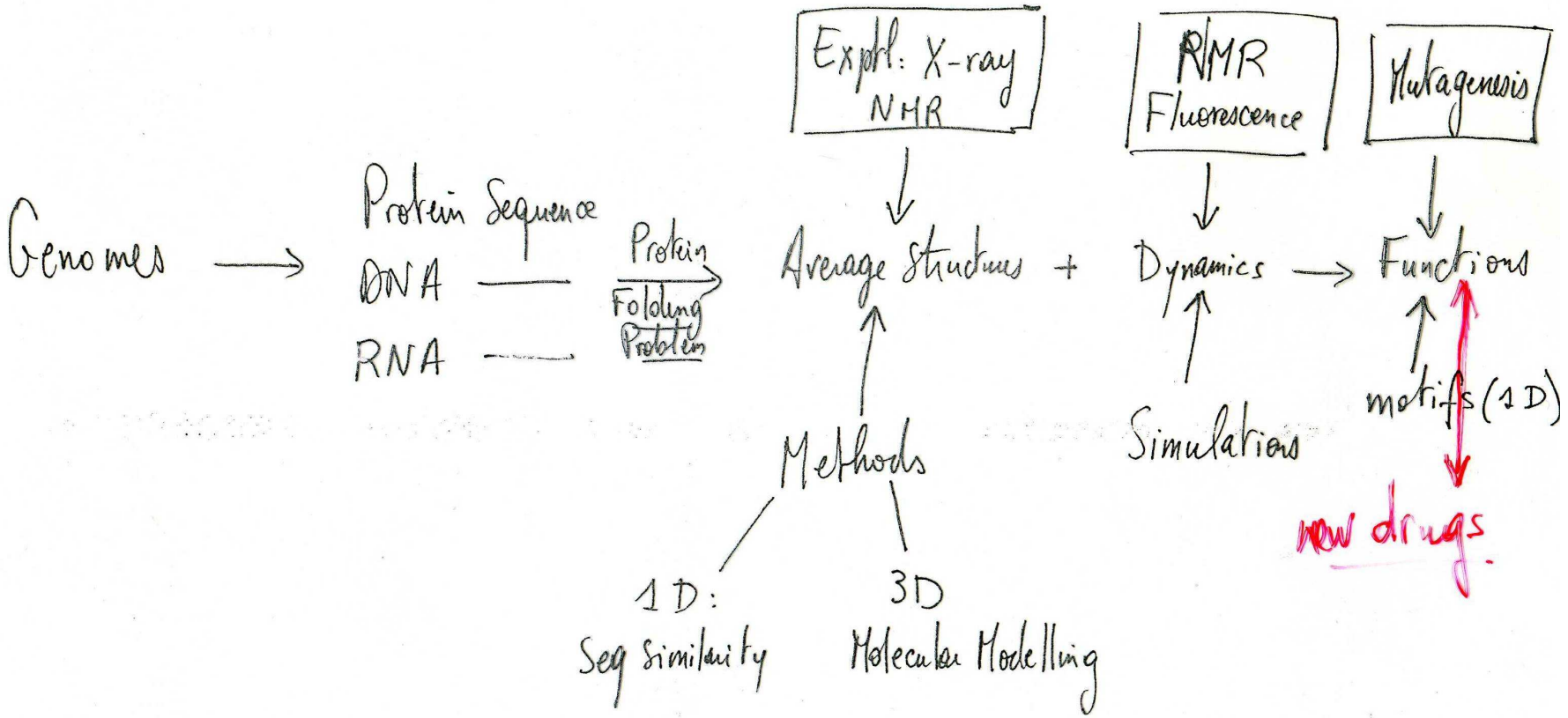
M1 Spécialité Bioinformatique Lecture 1A

P. Derreumaux

The protein folding problem: Generalities

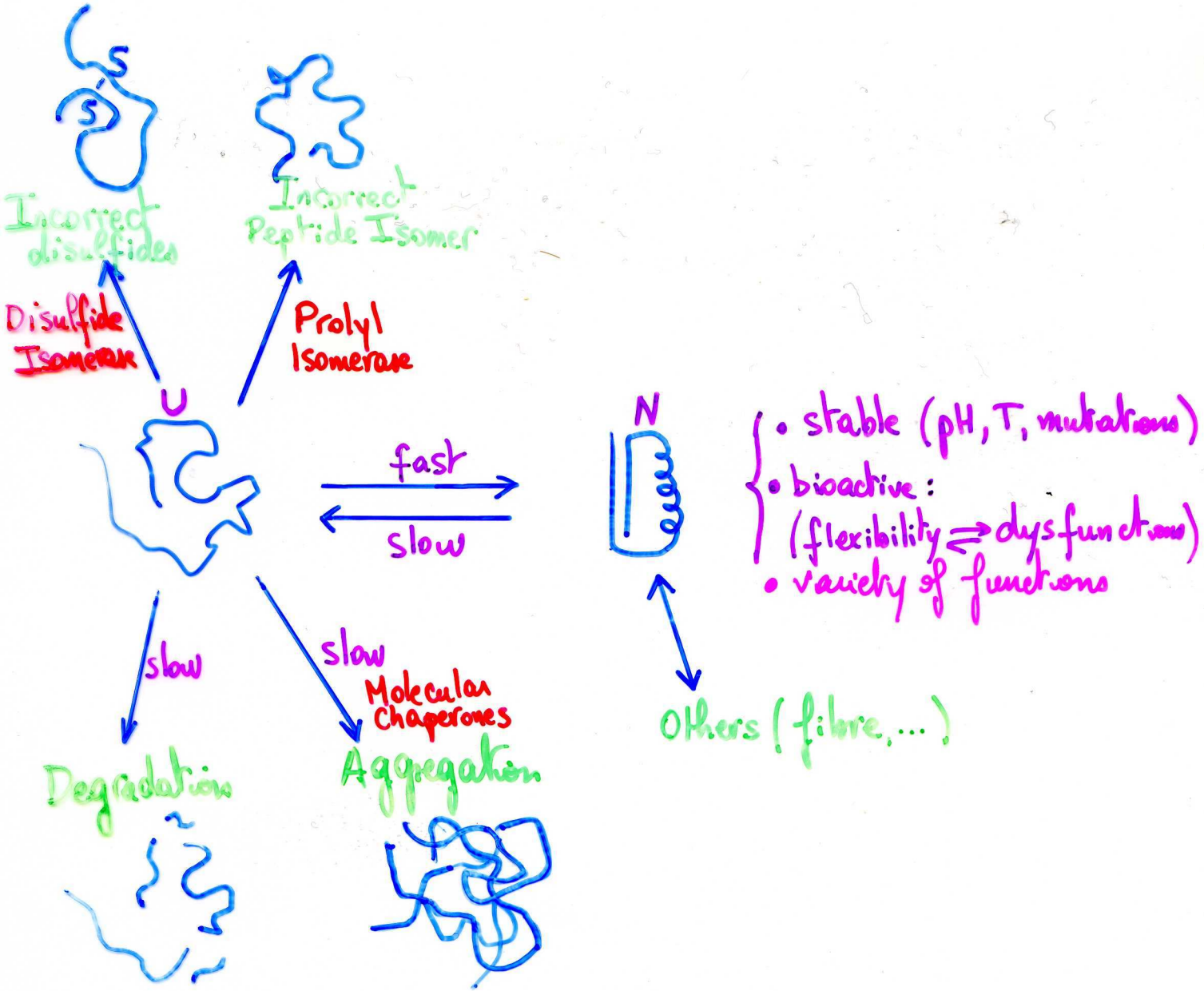

Genomes
Protein Sequence
DNA
RNA
Protein Folding Problem
Exptl: X-ray NHR
Average Structures +
Methods
1D: Seq Similarity
3D Molecular Modelling
NMR Fluorescence
Dynamics
Simulations
Mutagenesis
Functions
motifs (1D)
new drugs

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

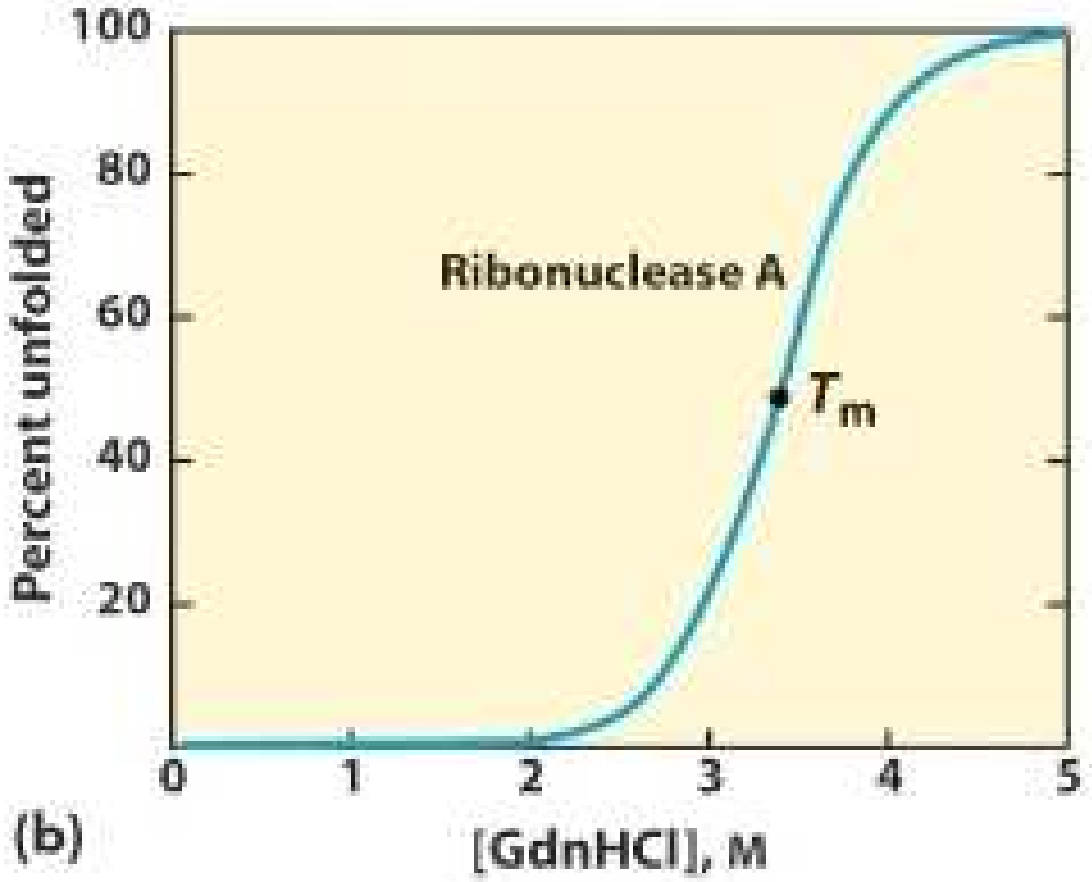

Incorrect disulfides
Incorrect Peptide Isomer
Disulfide Isomerase
Prolyl Isomerase
U
fast
slow
N
• stable (pH, T, mutations)
• bioactive :
(flexibility ⇌ dysfunctions)
• variety of functions
slow
slow
Molecular Chaperones
Others (fibers, ...)
Degradation
Aggregation

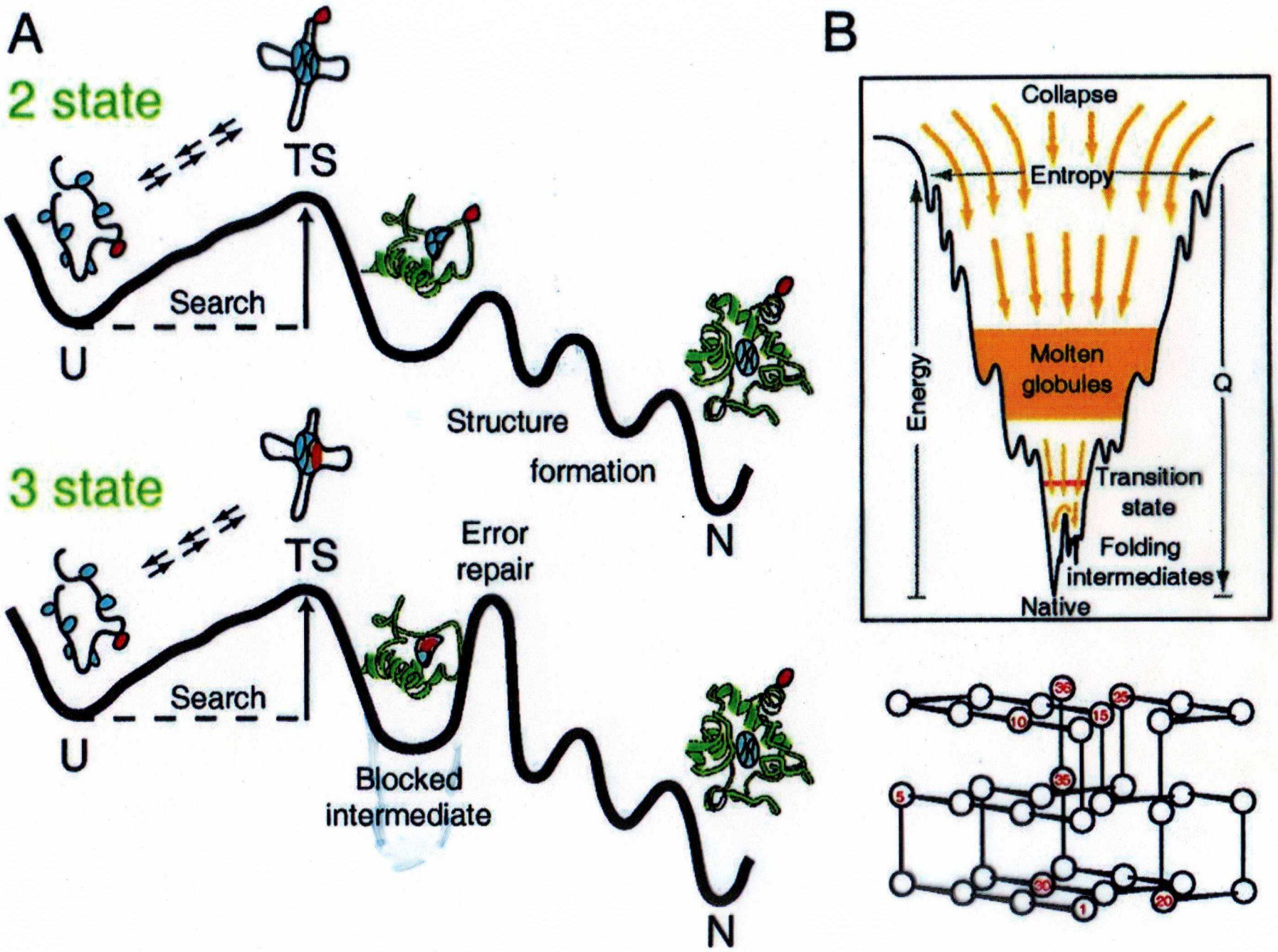
-Folding under chaperon control

-Thermodynamic hypothesis of folding (Anfinsen)

- Amyloid Fibril Formation

Monomeric Protein Folding: a simple experimental view




A
2 state
TS
Search
U
Structure
formation
N
3 state
TS
Error
repair
Search
U
Blocked
intermediate
N
B
Collapse
Entropy
Energy
Molten
globules
Q
Transition
state
Folding
intermediates
Native
36
25
10
15
35
5
30
1
20

Monomeric Protein Folding: a simple view

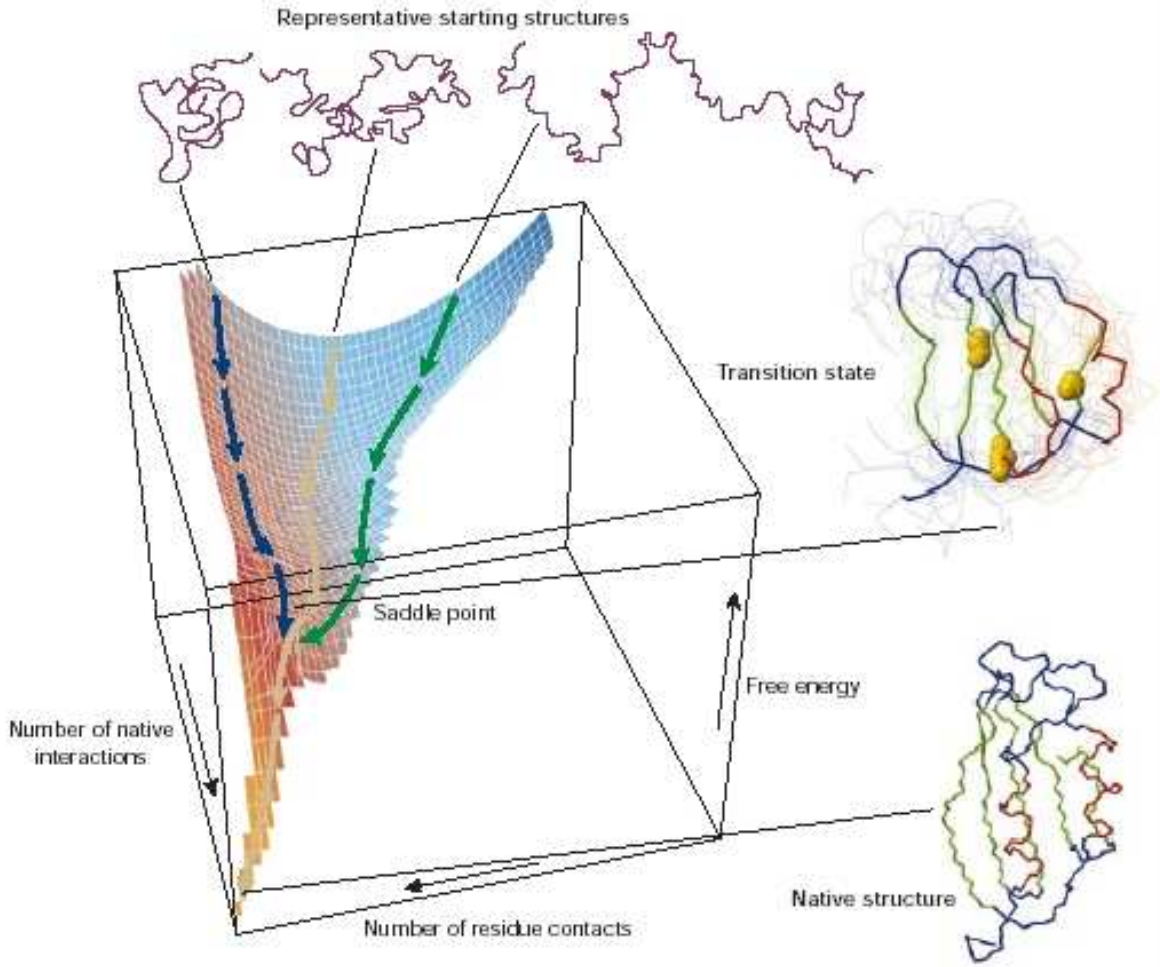
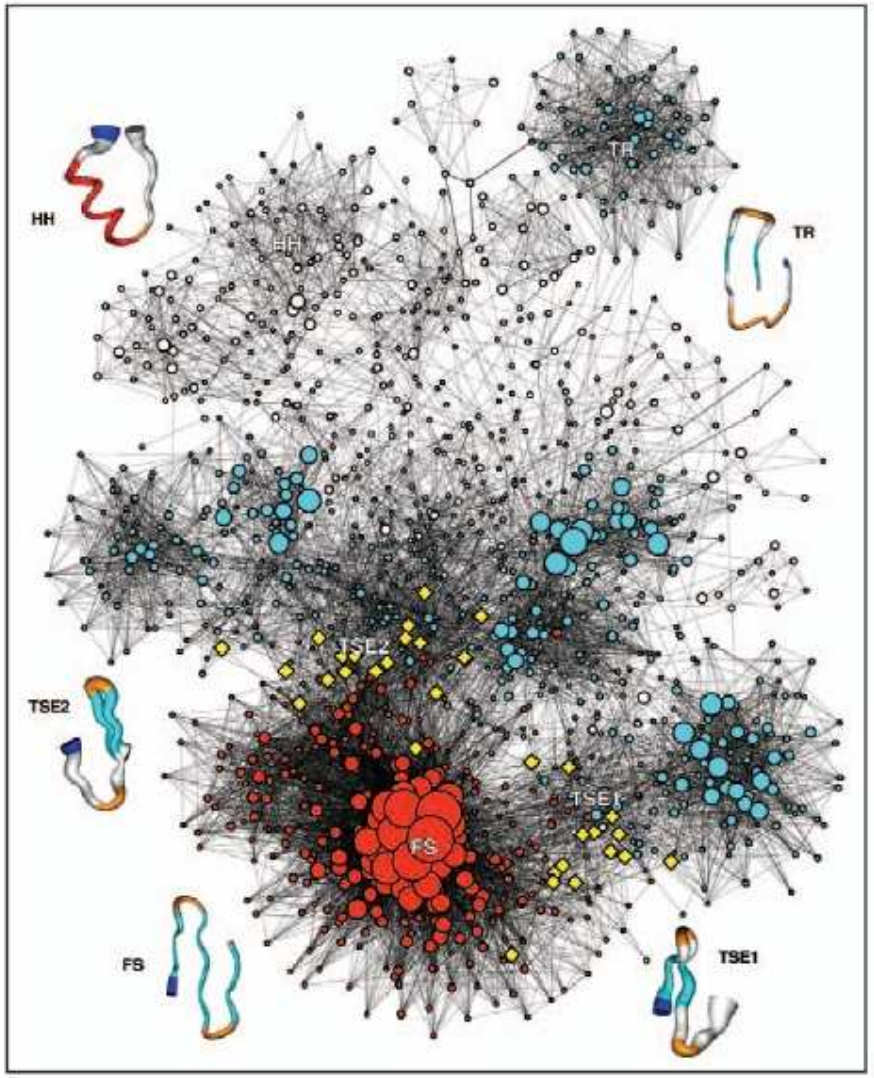


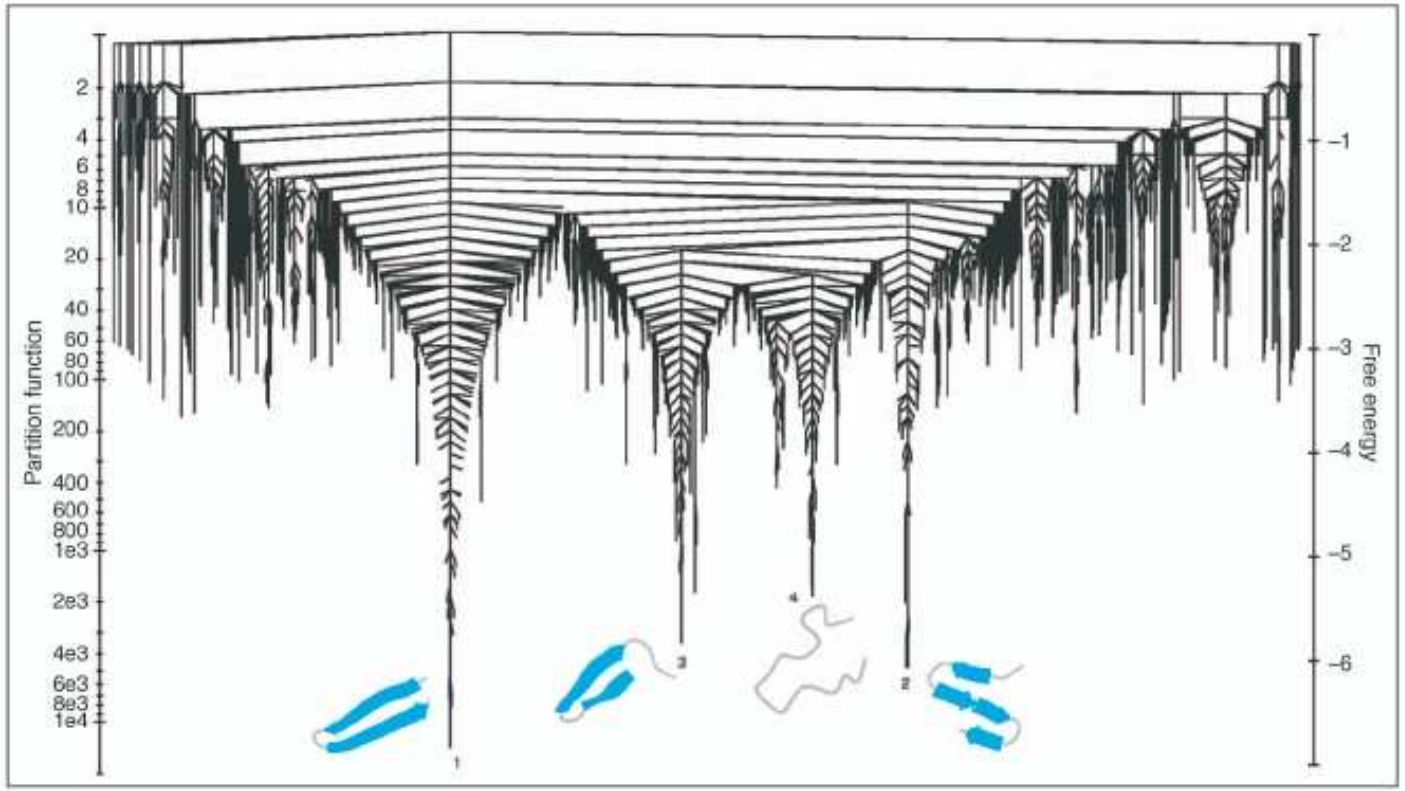
Figure 1 A schematic energy landscape for protein folding. The surface is derived from a computer simulation of the folding of a highly simplified model of a small protein. The surface 'funnels' the multitude of denatured conformations to the unique native structure. The critical region on a simple surface 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

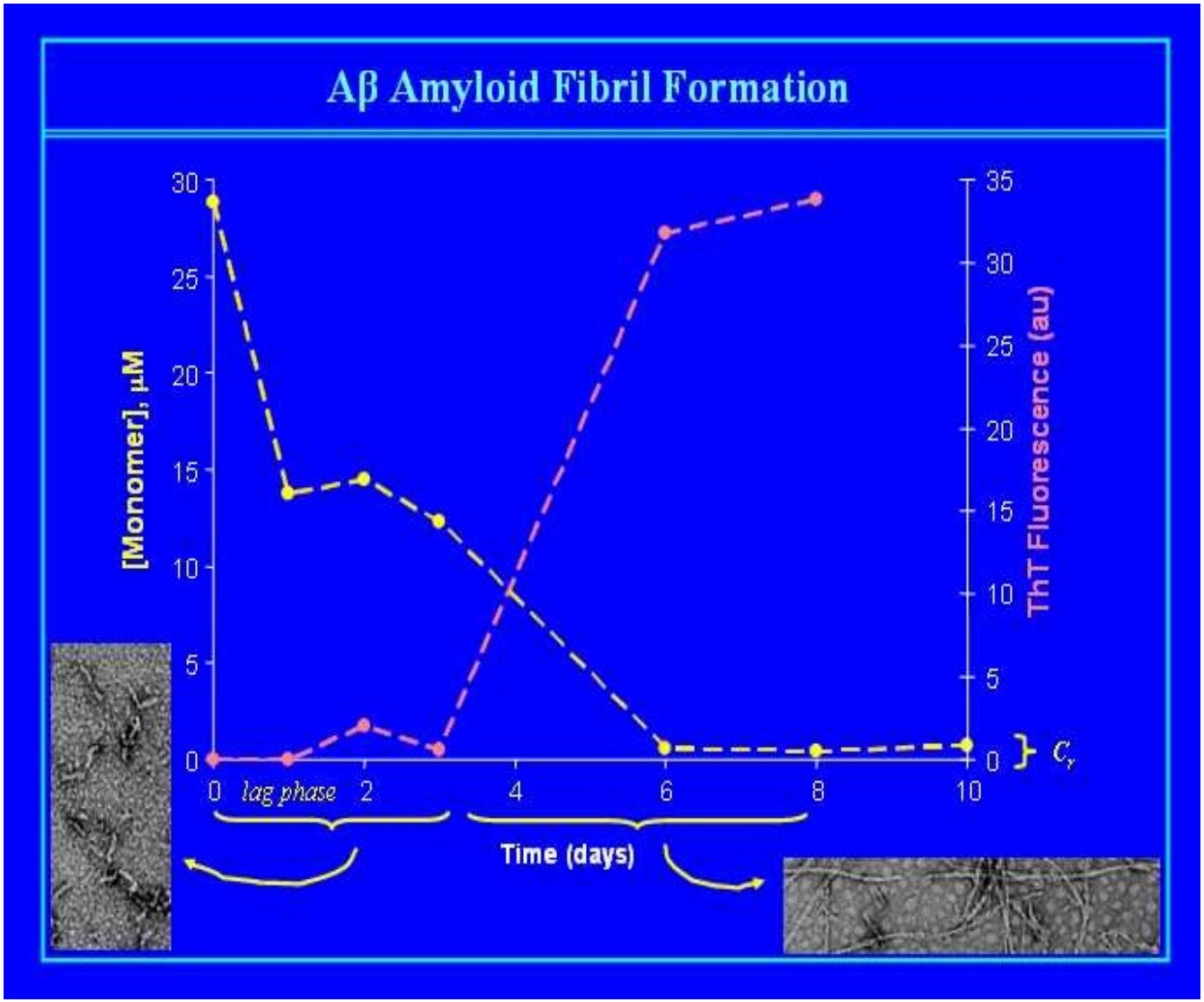


The funnel energy surface is oversimplified

Figure 3

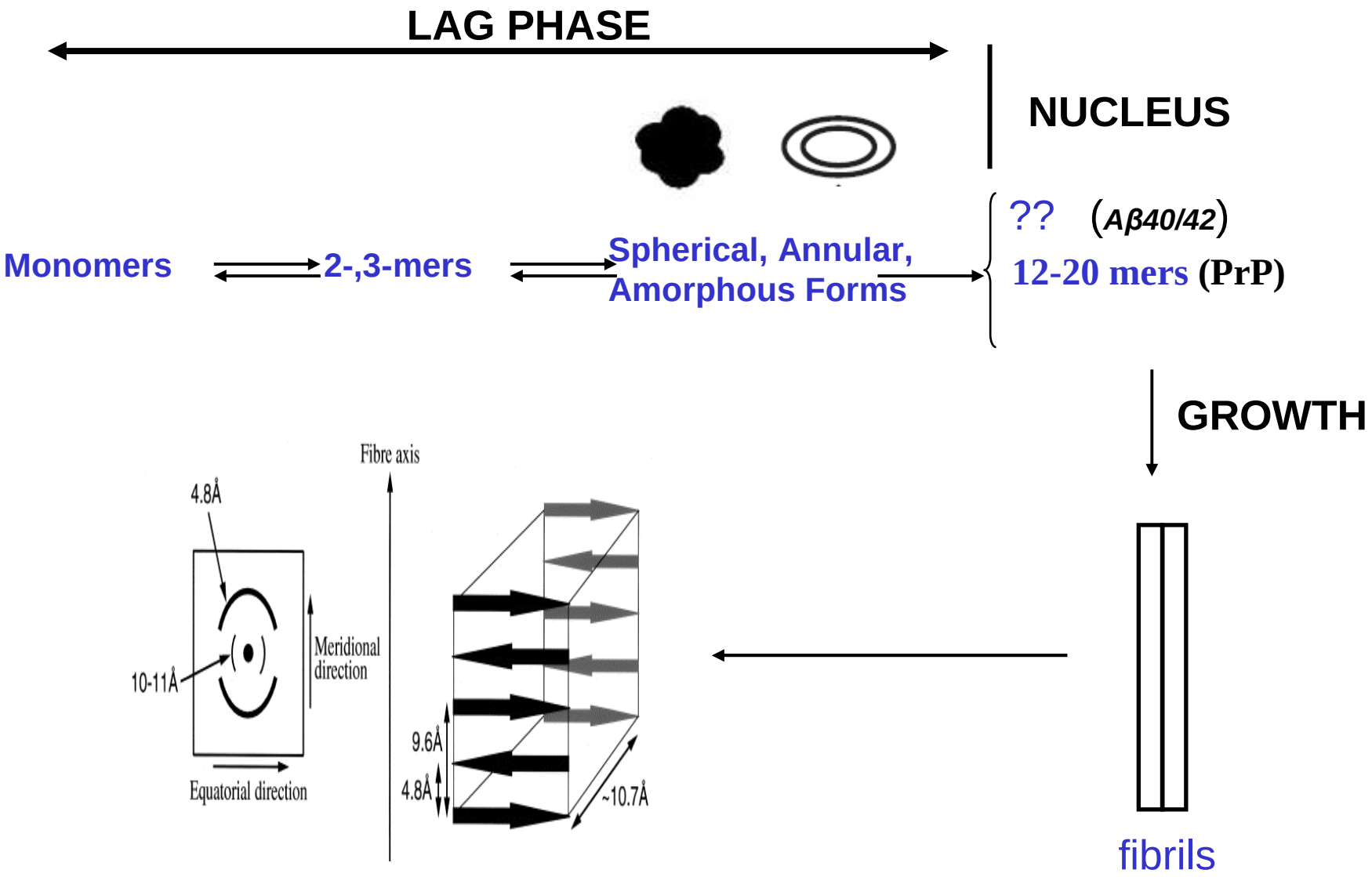


Transition disconnectivity graph of a β-hairpin (the C-terminal segment from the B1 domain of protein G). A total of 4 μs implicit solvent molecular dynamics simulations at 360 K were sampled to obtain a sufficient number of folding-unfolding events [29]. Representative structures of the deepest free 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].



Amyloid fibril formation is characterized by a polymerization-nucleation process

Details of the polymerization-nucleation process



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

Prediction 3D structure from sequence

-Genomic Programs and the number of sequences

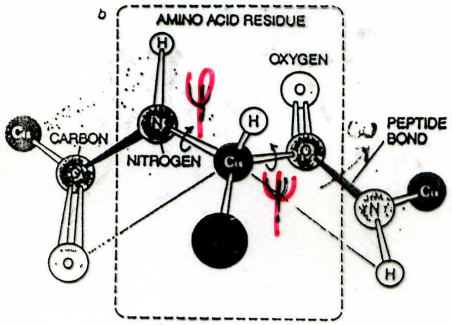
-Structures are more conserved than sequences

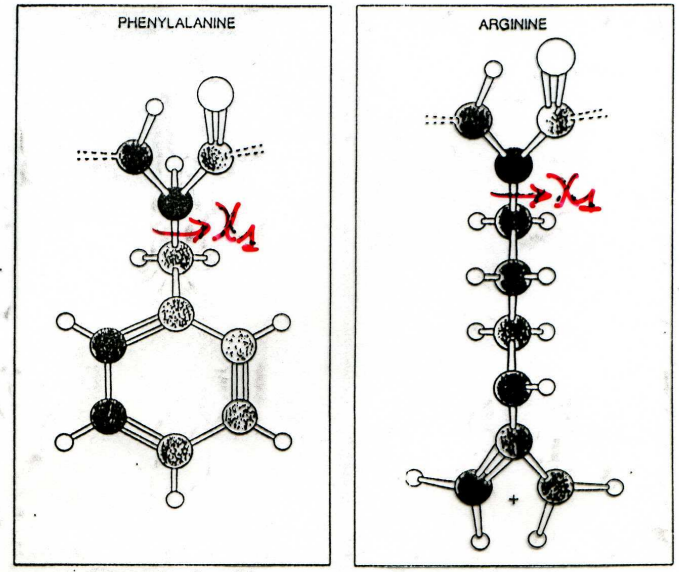
-Experimental costs for structure determination

A small number of sequences associated with functions

Alphabet 20 AA.

$N = 200$ AA $\rightarrow$ 20^{200} seq $>>>$ 10^{23}





DIFFERENCES in the shape, size and polarity of amino acids derive from differences in their side chains. In phenylalanine, for example, the side chain is nonpolar and cyclic, whereas the side chain of arginine is both strongly polar and linear.

A small number of 3D architectures for water-soluble proteins, $\simeq$ 450 known vs. $10^3 - 5\,10^3$ estimated

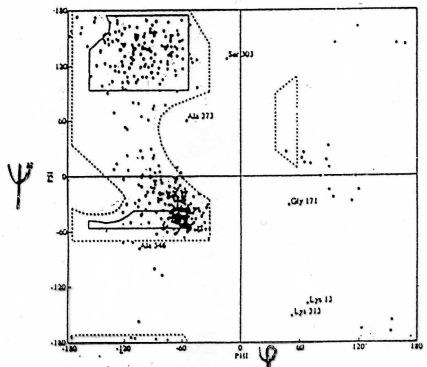
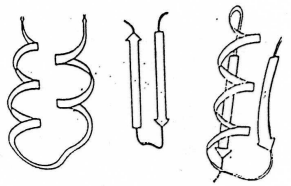
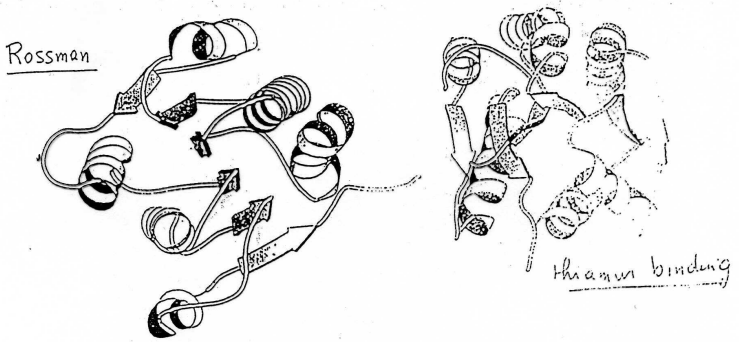
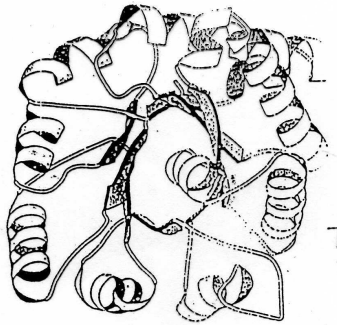


Fig. 1. Ramachandran plot of the refined hTIM structure. Open circles denote glycine residues, and the prolyl residues are indicated in triangles. Residues lying outside the allowed limits are marked (see text for details).







domain residues 213 to 326; Ammon *et al.*, 1968); (g) immunoglobulin, 3hhr (chain B, domain residues 32 to 131; De Vos *et al.*, 1992); (h) UB, 1sha (chain A; Waksman *et al.*, 1992); (i) jelly roll, 2stv (Fridborg *et al.*, 1965; (j) plaitfold, 1aps (Saudek *et al.*, 1989).

Table 1. Summary of average % of secondary structure residues which are in supersecondary structural units and motifs in all the categories of domain structure

Fold	*n*	% Secondary structure residues in: ββ, αα, βαβ	+(βαββ)	(β_4)
Globin	4	88.0	88.0	88.0
UpDown	6	90.3	90.3	90.3
Trefoil	4	83.4	83.4	90.2
TIM barrel	13	82.4	83.4	83.4
OB folds	5	76.7	76.7	81.4
Doubly wound	38	68.1	70.3	70.3
Ig	17	66.7	66.7	85.6
UB roll	6	55.3	55.3	61.1
Jelly roll	6	47.3	47.3	67.9
Plaitfold	23	37.8	62.8	64.0
All superfolds	122	64.7	70.5	74.8
Non-superfolds	516	61.7	63.4	66.3

The *n* value is the number of domains per fold. The ββ, αα, βαβ column is the mean value for % of secondary structure residues in supersecondary structural units. The +(βαββ) column is the mean value for % of secondary structure residues in supersecondary structural units and in the βα-Greek key motif. The +β_4 column is the mean value for % of secondary structure residues in supersecondary structural units and in the βα-Greek key and in the β_4 Greek key motifs. This is the order in which the motif additions were made to the data.

PDB 1999 ≃ 50% secondary structures

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

basic reasons:

obvious
- 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)

not obvious
- 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.

IDP: Intrinsincally Disordered Proteins

- 30% of eukaryotic genome encoded proteins

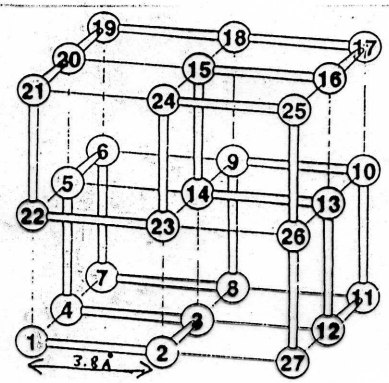
- disordered over the full or part of the sequence

- fold upon binding on one or multiple partners

Thermodynamic, Kinetic and dynamic aspects

Mostly lattice Models.
(C_α)

Off-lattice C_α Models (Nymeyer, PNAS, 95, 1998)



Thermodynamic Properties
full enumeration $N_{c,GEM}^{*}$ known
Canonical Ensemble

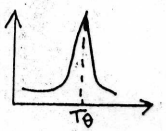
$$Q = \sum_{i}^{N_c} e^{-\frac{E_i}{k_B T}}$$

↳ $A = -k_B T \ln Q = E - TS$

↳ $P_{Nat} = \frac{e^{-\frac{E_{N_{AT}}}{k_B T}}}{Q} \rightarrow T_f$

↳ $C_V \rightarrow T_\theta$

↳ $S \rightarrow T_g$ $\left(\frac{\partial S}{\partial T} = 0\right)$



$N_c = N$ (30-mers 2D, 15-mers 3D)
$= N_{compact}$ (27-mers, 3D)

Kinetic Properties
run MC Metropolis
- $\min\left(1, e^{-\frac{\Delta E}{k_B T}}\right)$
- moves

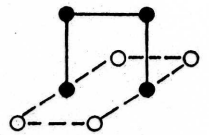


Figure 3. Crankshaft motion. The original
is indicated by solid lines and filled circles.
conformations are indicated by dashed line

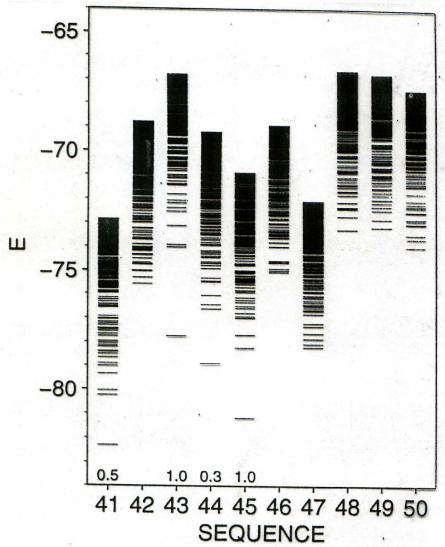
- $\pi(t) = f(t)$ in function of sequence, T, PEF

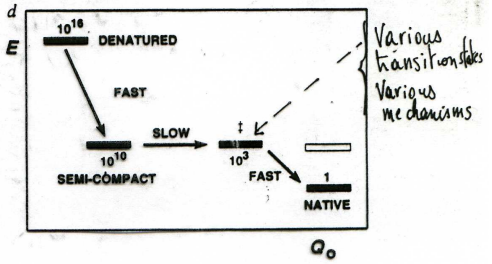
What have we learned ?

- Small Peptide model (3x3x3, 4x4x4)

$$Z = \frac{E_{NAT} - E_U}{\sigma_U} \rightleftharpoons \text{fast folding.}$$

(Sali, Nature, 369, 1994)





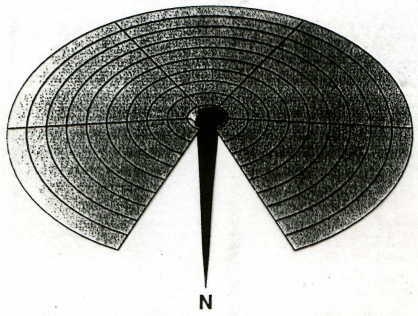
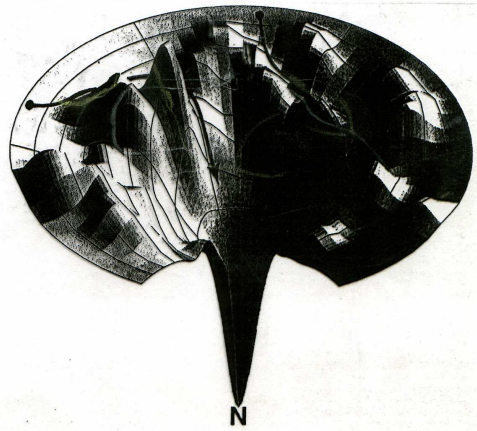


Fig. 1 The Levinthal 'golf-course' landscape. **N** is the native conformation. The chain searches for **N** randomly, that is, on a level playing field of energies.



(Dill, Nat. Struct. Biol., 4, 1997)

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

M1 Spécialité Bioinformatique Lecture 1B

P. Derreumaux

Predicting Protein Structures from AA:

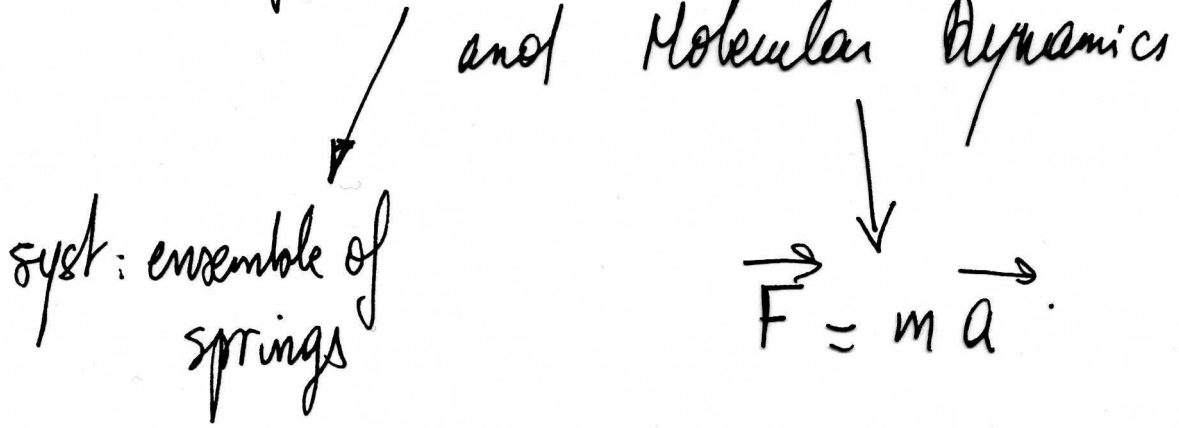
Two perspectives

1. If you were a theoretical chemist or physicist
2. If you were a pure bioinformatician

From Quantum Mechanics.

$$H\psi = E\psi$$

to Molecular Mechanics and Molecular Dynamics



Niveau d'approximation en Mécanique classique

Born-Oppenheimer : les mouvements des électrons sont découplés des mouvements des noyaux.

Les électrons sont représentés par leurs effets : charges partielles, paramètres (distance de référence entre 2 atomes liés de façon covalente, ...).

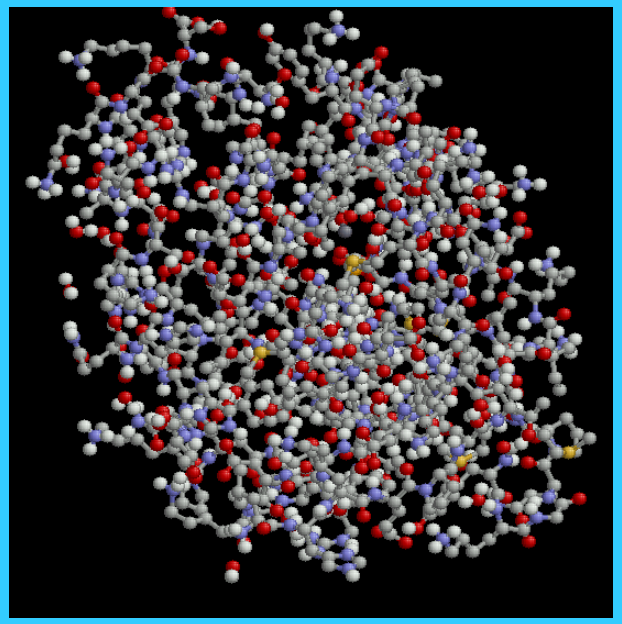
Seuls, les mouvements des noyaux sont considérés et sont traités dans un modèle de ***mécanique classique***.

→ champs de forces **semi-empiriques**
→ grands systèmes, typiquement < 10000 atomes
→ solutions numériques
→ dynamique, limitée à ~ 1 ns (10^{-9} s)
→ représentation du solvant et des contre-ions
→ énergie conformationnelle
→ pas de réactions chimiques, polarisation, transfert d'électrons

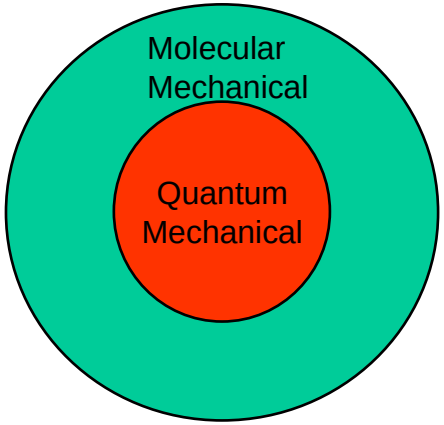
Capacité de transfert : des paramètres calculés et testés sur de petites molécules peuvent être utilisés pour des systèmes plus importants.

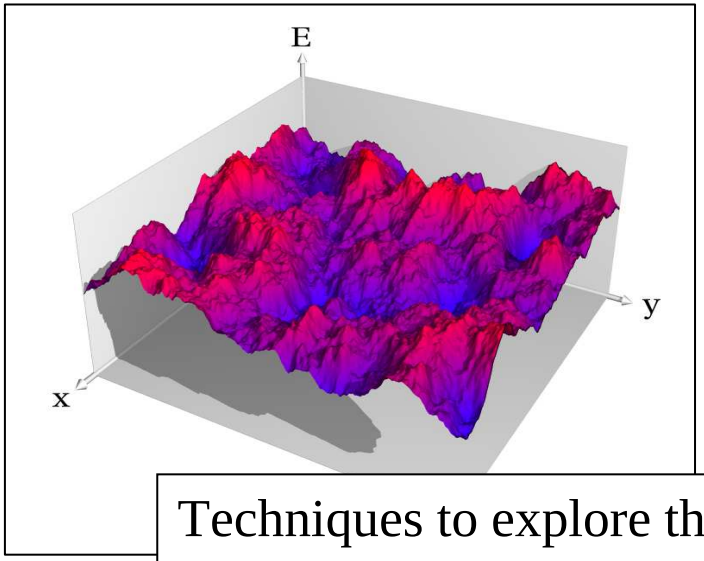
Computer Simulation - Basic Principles

Model System



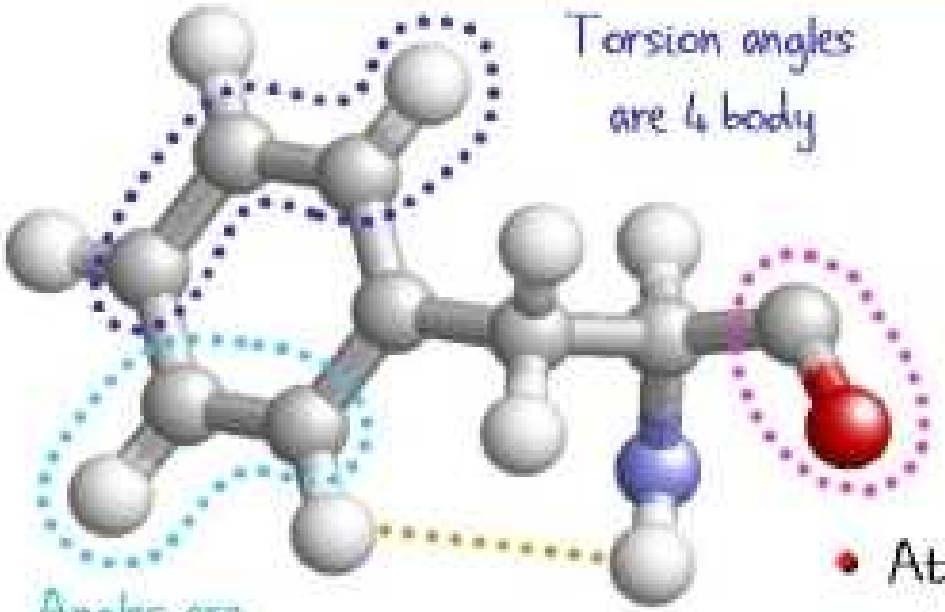
Coarse-grained
Force field





Techniques to explore the
energy landscape
(structures, thermodynamics)

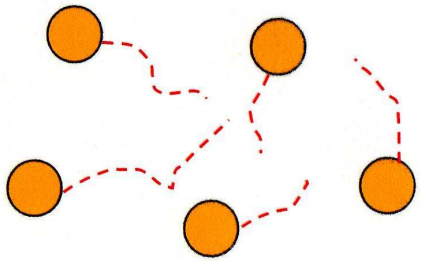
WHAT IS A MOLECULE?



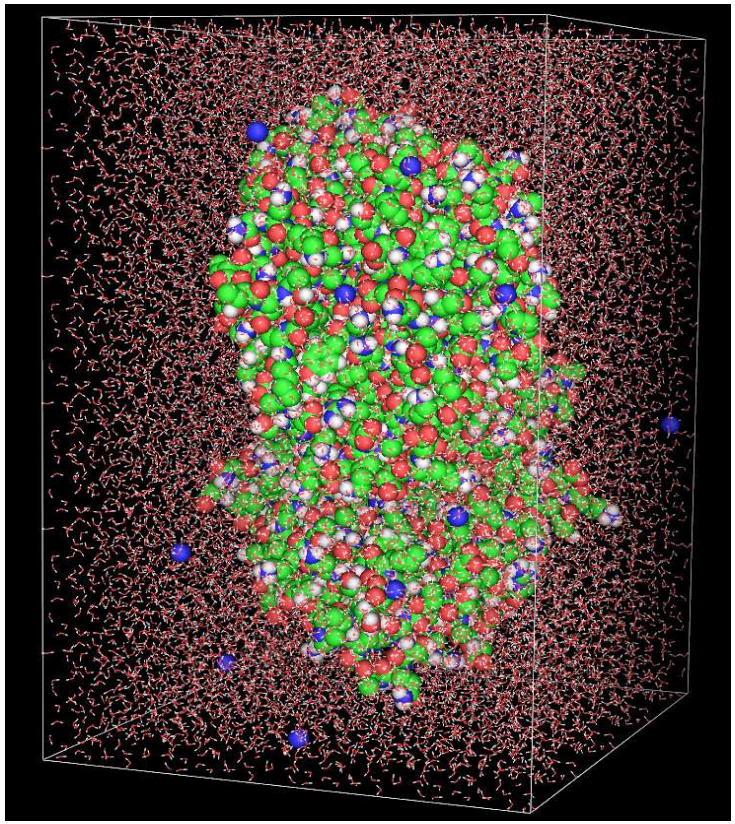
- Atoms are connected in a simple graph.
- Bonded Interactions: 2, 3, 4 body.
- Non-Bonded Interactions.

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.



Treatment solvent



Box or

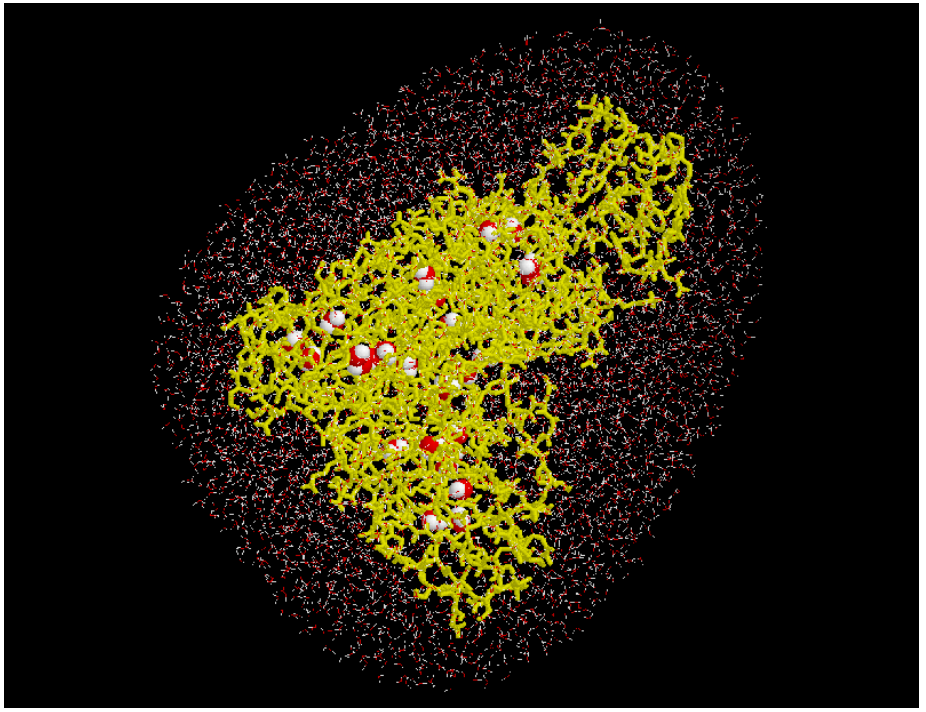
Droplet ?

SPC, TIP3P..?

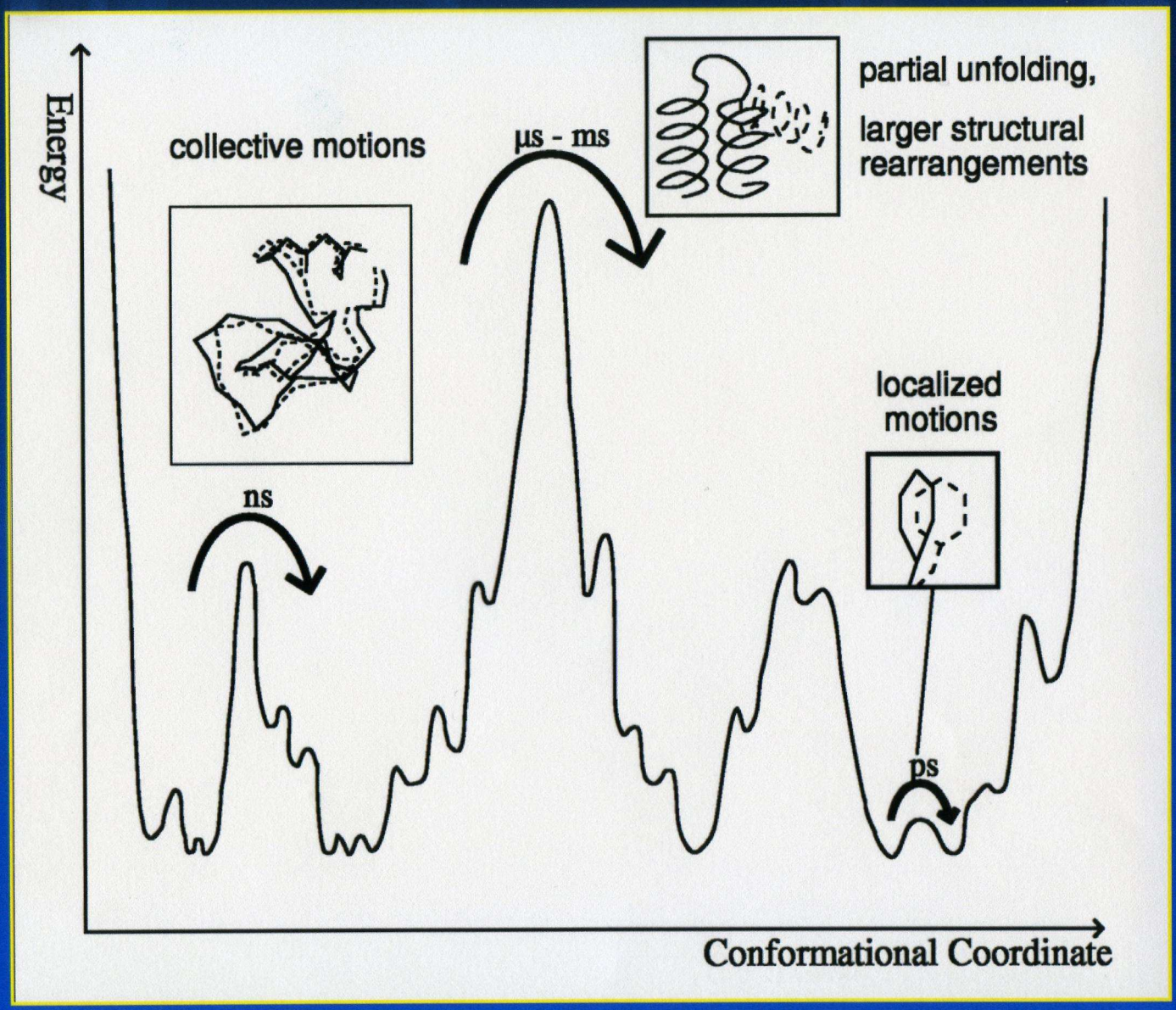
explicit

or

implicit?



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



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

Protein Folding: Fast Folders

Time Scale:

- 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

Ideally : MD simulations in solvent.

MD simulation : a simple application.

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

Suff $\quad a = \dot{v} = \frac{dv}{dt}$

$$v = at + v_0$$

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

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

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

$$\boldsymbol{v}(t + \frac{\Delta t}{2}) = \boldsymbol{v}(t - \frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m}\Delta t$$

$$\boldsymbol{r}(t + \Delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t + \frac{\Delta t}{2})\Delta t$$

with

$\Delta t \approx 1fs!$

(equivalent to the Verlet algorithm)

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

Molecular dynamics

36 amino acids
+3,000 water
molecules

1 ns simulation
(expt'l 5-10 μs)

256 parallel processors
on a CRAY T3E

2 months of computer
time

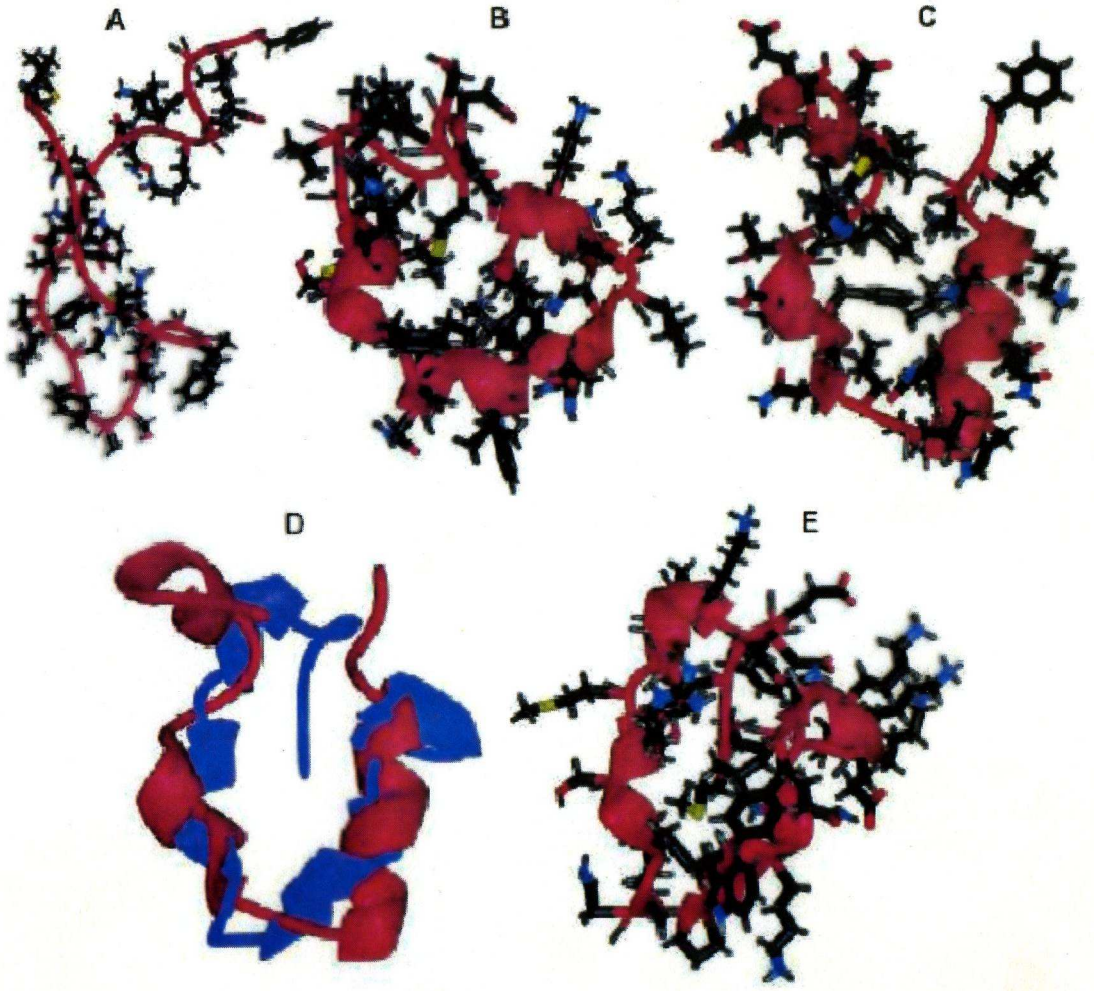


Fig. 1. Ribbon representations of (A) the unfolded, (B) partially folded (at 980 ns), and (C) native structures, and (E) a representative structure of the most stable cluster and (D) the overlap of the native (red) and the most stable cluster (blue) structures, generated with UCSF MidasPlus. Color code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

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

Atomic-Level Characterization of the Structural Dynamics of Proteins

David E. Shaw,[1,2]* Paul Maragakis,[1]† Kresten Lindorff-Larsen,[1]† Stefano Piana,[1]† Ron O. Dror,[1] Michael P. Eastwood,[1] Joseph A. Bank,[1] John M. Jumper,[1] John K. Salmon,[1] Yibing Shan,[1] Willy Wriggers[1]

Molecular dynamics (MD) simulations are widely used to study protein motions at an atomic level of detail, but they have been limited to time scales shorter than those of many biologically critical conformational changes. We examined two fundamental processes in protein dynamics—protein folding and conformational change within the folded state—by means of extremely long all-atom MD simulations conducted on a special-purpose machine. Equilibrium simulations of a WW protein domain captured multiple folding and unfolding events that consistently follow a well-defined folding pathway; separate simulations of the protein's constituent substructures shed light on possible determinants of this pathway. A 1-millisecond simulation of the folded protein BPTI reveals a small number of structurally distinct conformational states whose reversible interconversion is slower than local relaxations within those states by a factor of more than 1000.

Many biological processes involve functionally important changes in the three-dimensional structures of proteins. Conformational changes associated with protein folding (*1*), signal transduction (*2*), the catalytic cycles of enzymes (*3*), and the operation of molecular machines and motor proteins (*4*) often involve transitions among two or more structur-

ally distinct st
terized as "ba
"energy landsc

Substantial
both experime
techniques, in c
and the ways th
them. It has pr
ally characteriz
states and to el
anisms involve
states.

All-atom n
tions are desig
view of the n
cules (*9*), prod
the potential t
shots generate
tational constr

[1]D. E. Shaw Rese
10036, USA. [2]Cen
matics, Columbia
*To whom corres
David.Shaw@DES
†These authors c

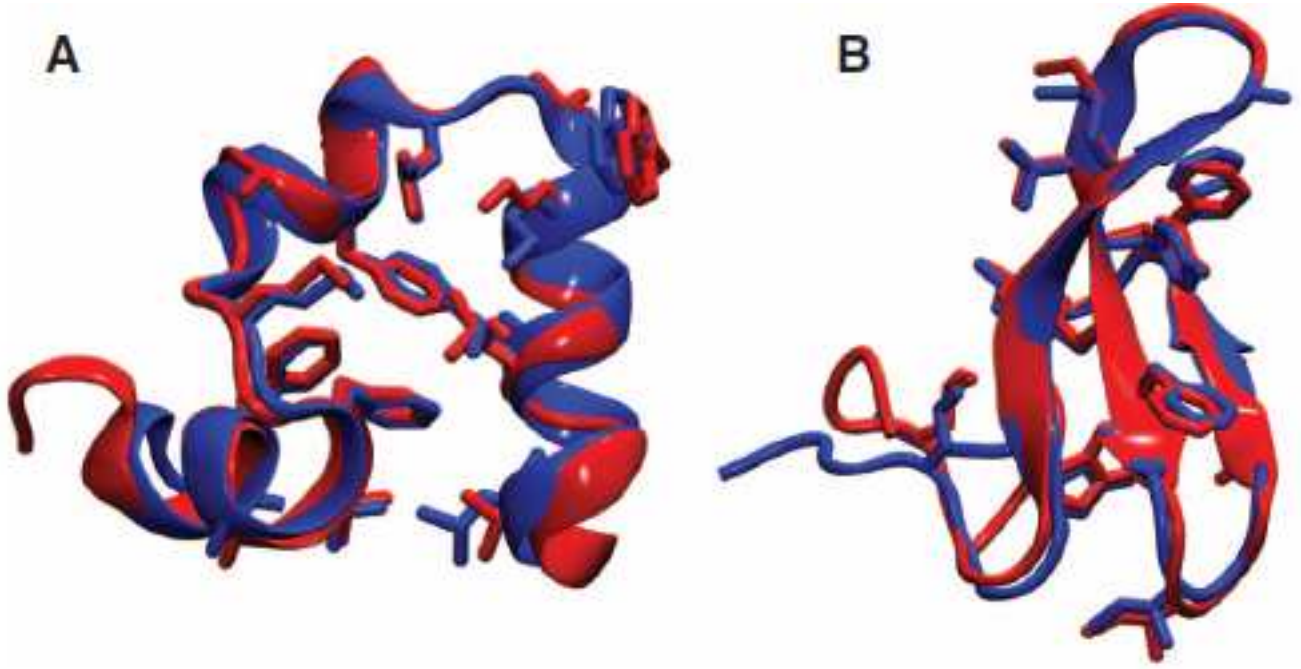
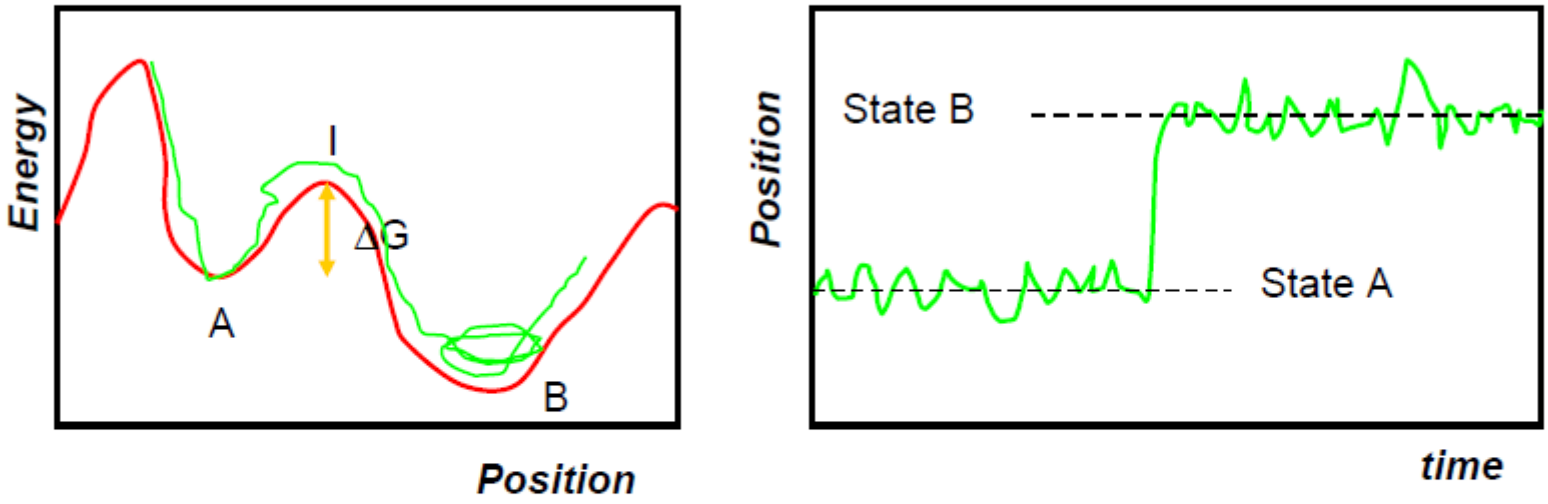


Fig. 1. Folding proteins at x-ray resolution, showing comparison of x-ray structures (blue) (*15*, *24*) and last frame of MD simulation (red): (**A**) simulation of villin at 300 K, (**B**) simulation of FiP35 at 337 K. Simulations were initiated from completely extended structures. Villin and FiP35 folded to their native states after 68 μs and 38 μs, respectively, and simulations were continued for an additional 20 μs after the folding event to verify the stability of the native fold.

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:

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