Curso del Master de Física:
Elementos de Física Biologica

Protein Folding: a Physical Perspective

Pierpaolo Bruscolini
Instituto BIFI – Universidad de Zaragoza
pier@unizar.es

- Proteins and protein folding, characteristics, time scales
- Levinthal paradox: how can we deal with it

- experimental facts
- Forces inducing folding

- How can we model protein folding?
- All-atoms models and empirical potentials
- Simple models: homopolymers, random bond heteropolymers

- Random energy model, with an eye on Kramers' rate theory

- folding funnel and energy landscapes: microscopic and macroscopic characterization
- HP models, many letters models
- Simple off-lattice models

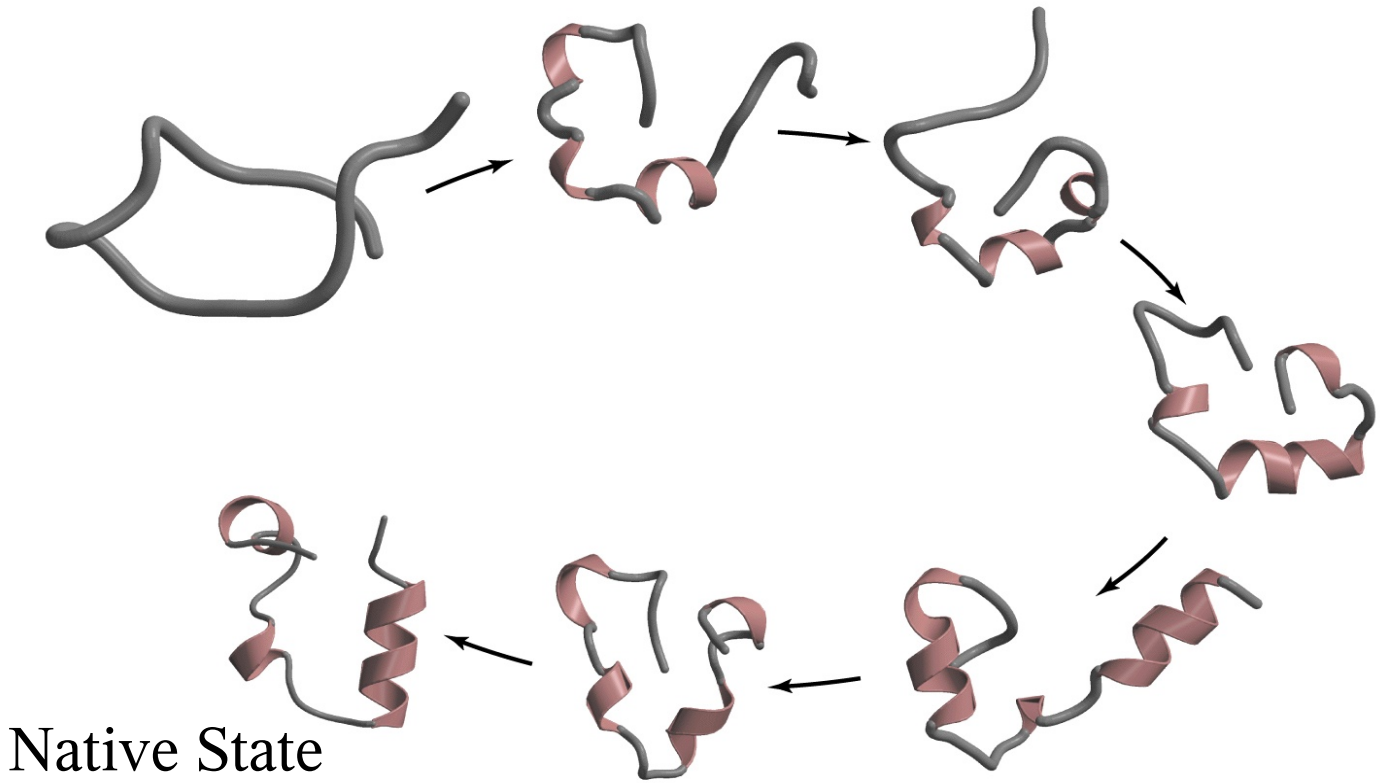
- Go models
- WSME model

HORATIO
O day and night, but this is wondrous strange!

HAMLET
And therefore as a stranger give it welcome.
There are more things in heaven and earth, Horatio,
Than are dreamt of in your philosophy.

... I'll try to show you the wonders of a strange phenomenon, whose existence would be difficult to imagine, if we didn't see it...

Protein Folding: the process by which a swollen protein reaches it functional, **Native State**.

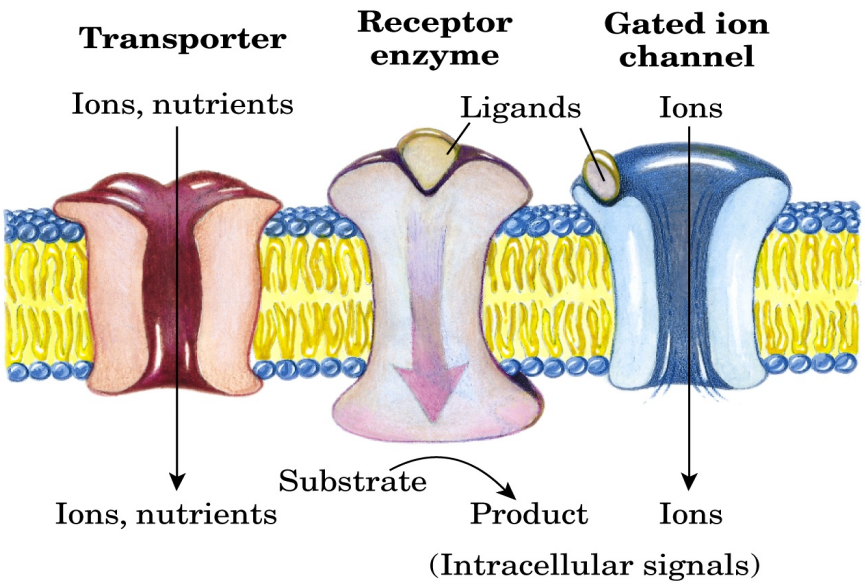


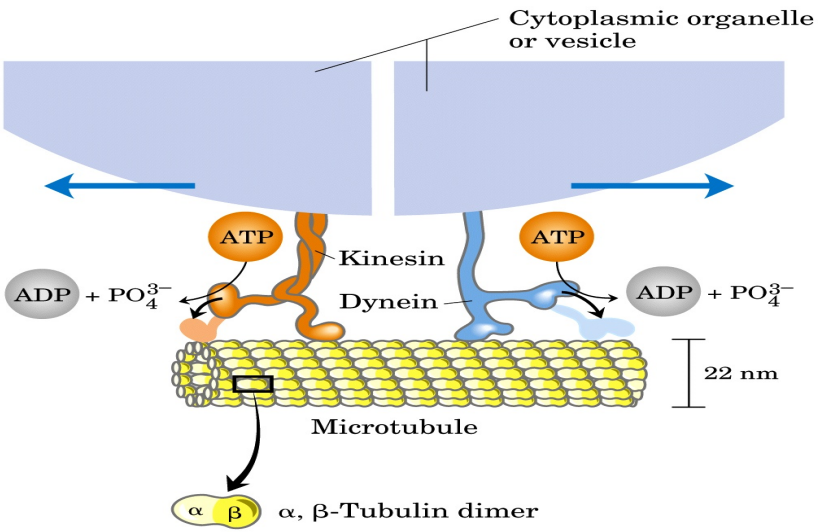
We will focus on **globular protein**, that fold in water. **Membrane proteins** live in a non-polar environment: they are very difficult to study experimentally.

PROTEINS

Linear Polymers, 50 ÷ >4000 amino acids (typically ~200 aa, ~3000 atoms)

Variety of biological functions (enzymatic, structural, transport, electron transport, movement, signal transduction...)

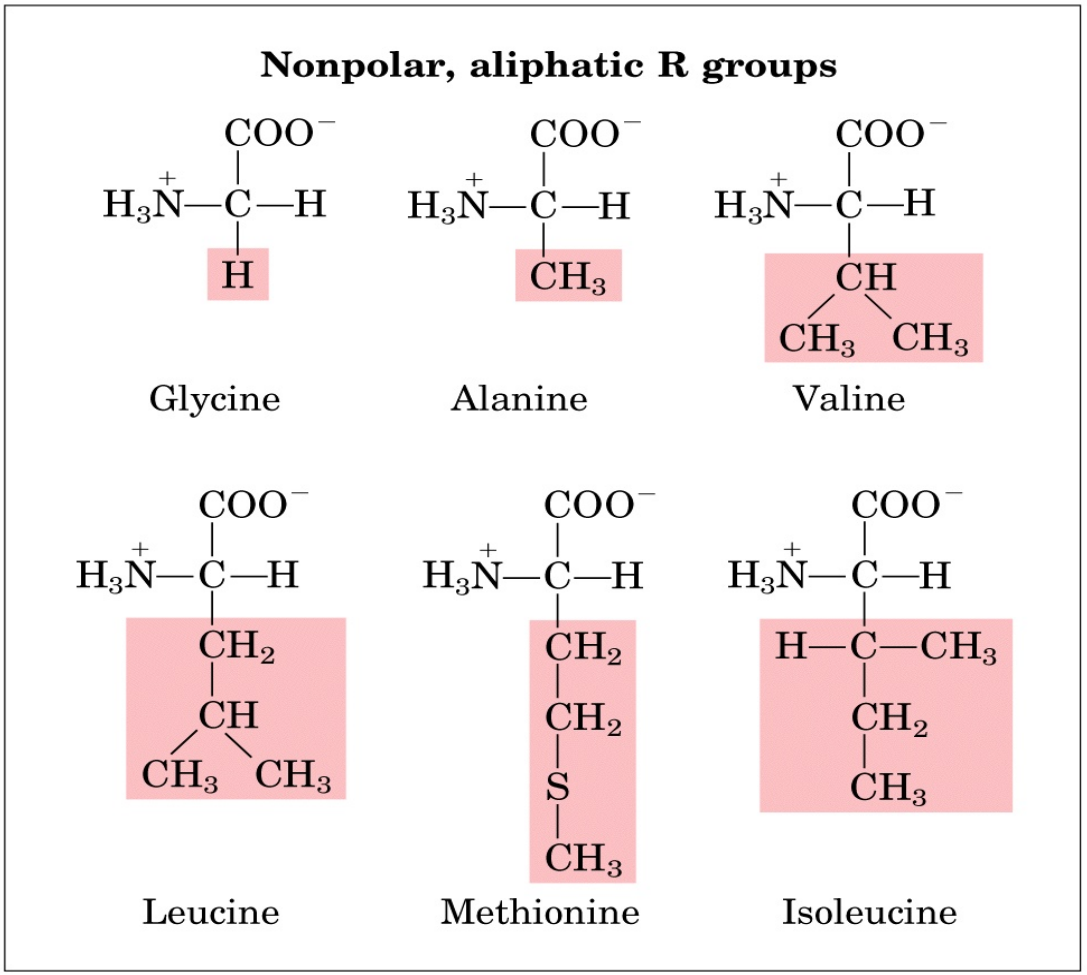




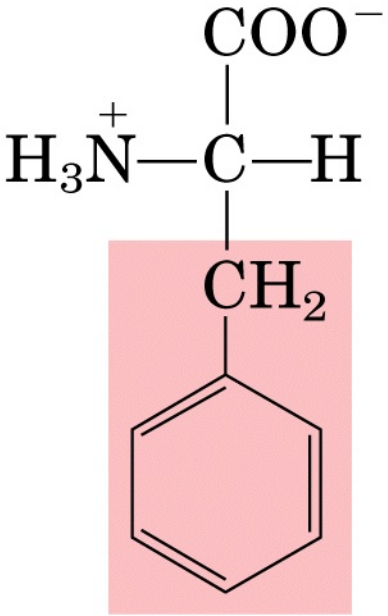
Proteins can function only when they are in a particular 3D conformation **(Native State)**

The building bricks: aminoacids (20 different kinds, with a common and a varying part (side chain))

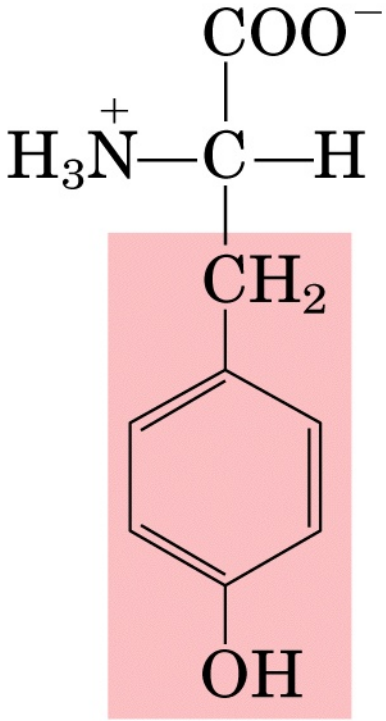
Non-polar (aliphatic, aromatic), Polar, Charged (positive, negative)



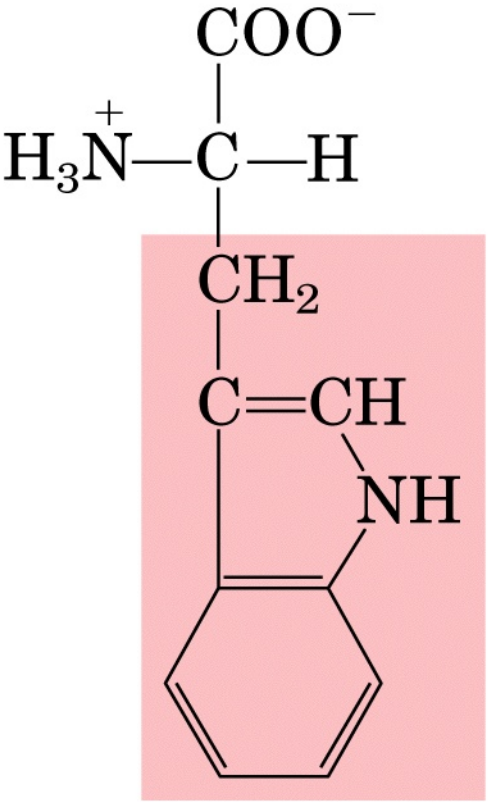
Aromatic R groups



Phenylalanine

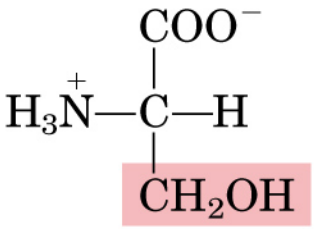


Tyrosine

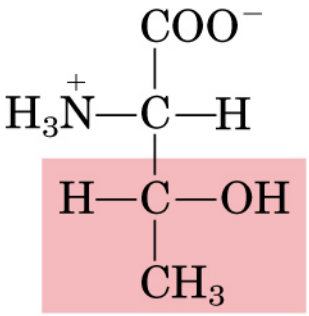


Tryptophan

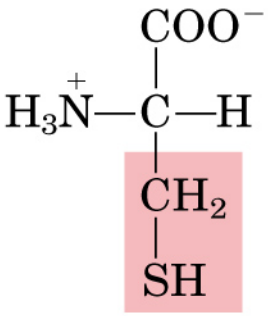
Polar, uncharged R groups



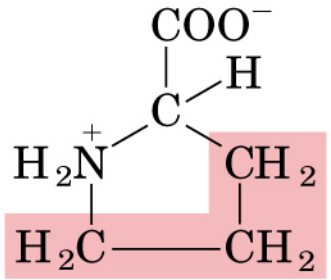
Serine



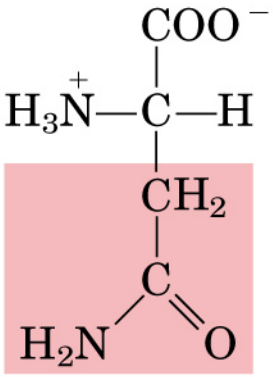
Threonine



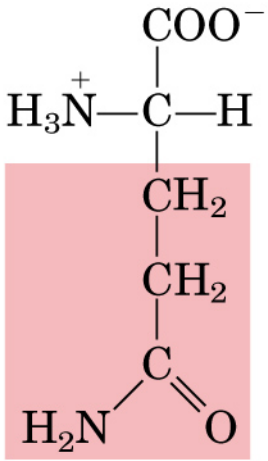
Cysteine



Proline

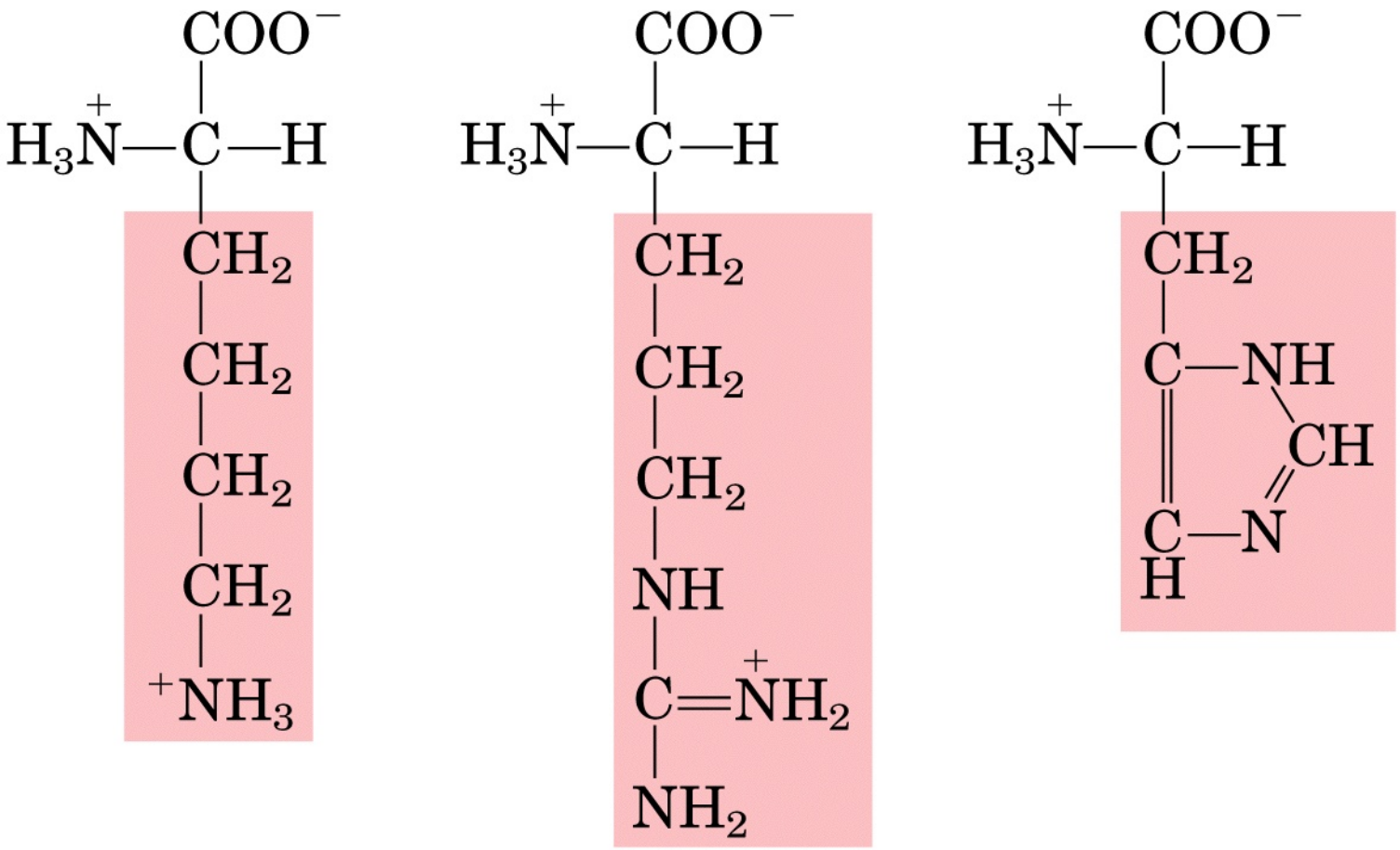


Asparagine



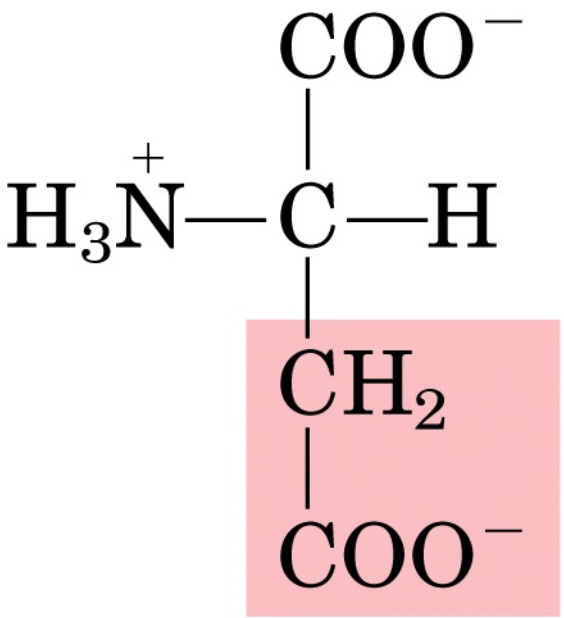
Glutamine

Positively charged R groups

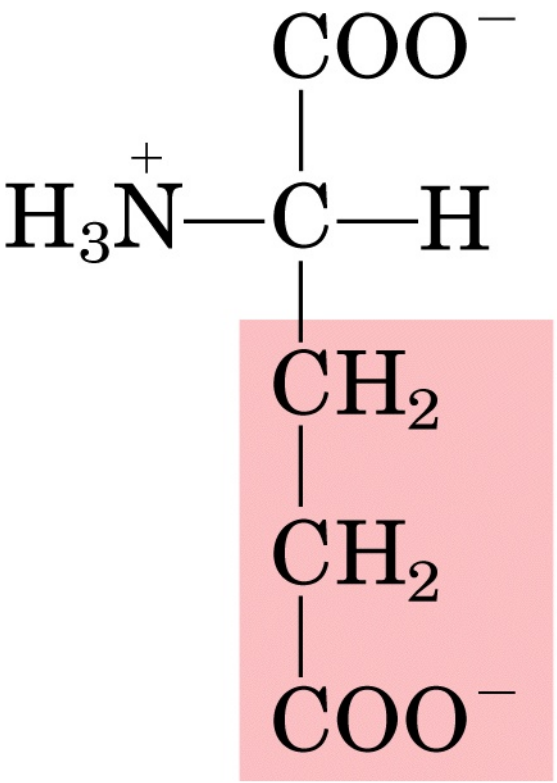


Lysine Arginine Histidine

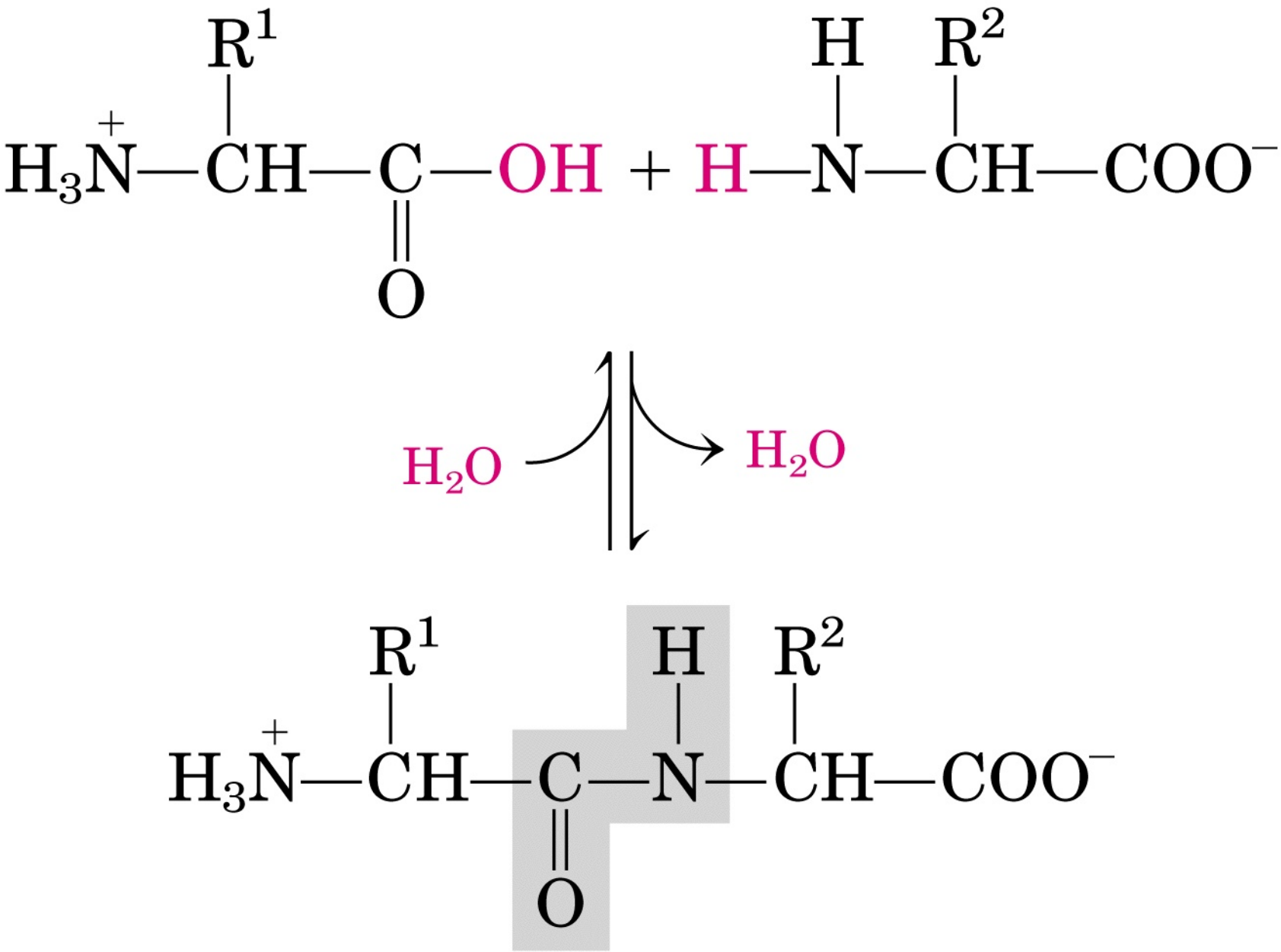
Negatively charged R groups



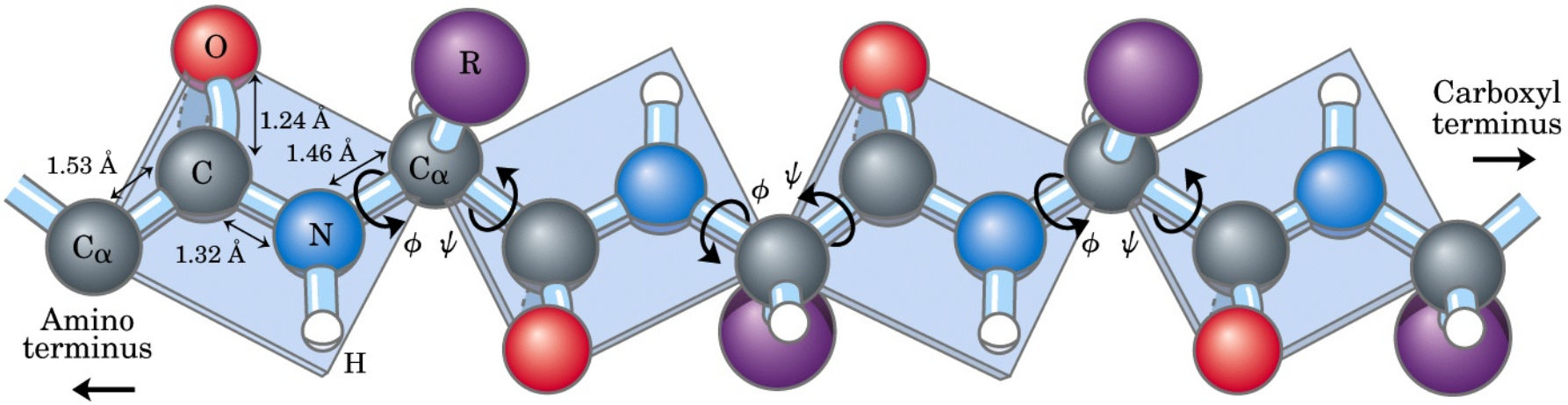
Aspartate



Glutamate


R1
H
R2
H3N+—CH—C—OH + H—N—CH—COO−
O
H2O
H2O
R1
H
R2
H3N+—CH—C—N—CH—COO−
O

Polymerization of the aminoacids: the "**primary structure**".
Aminoacids in proteins are called "residues".

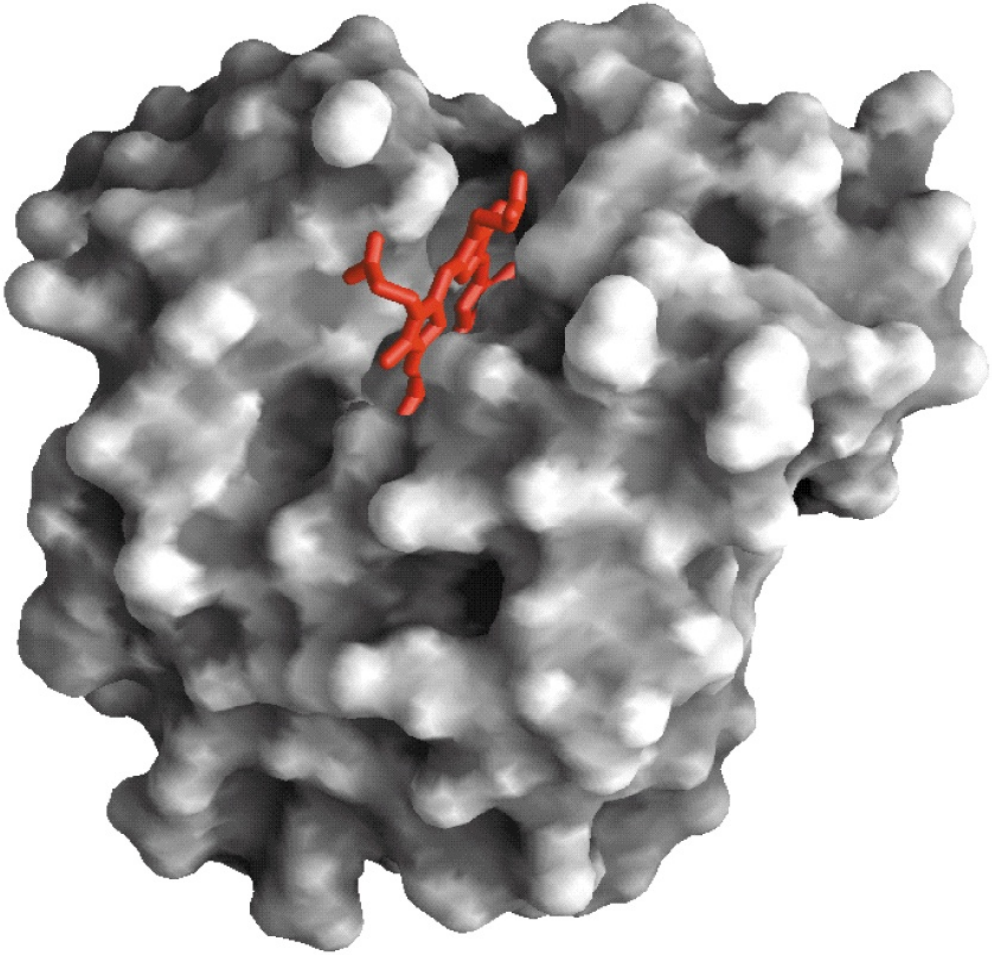


(b)

R = side chain (20 types)
Peptide units (the planes in the figure) make up the **backbone**, or **main chain**.

The primary structure involves (highly energetic) covalent bonds: it remains fixed during protein's life. Hence the **sequence of residues is fixed, and different for each protein**. For a protein of 200 a.a., 20^{200} sequences are possible... (*are they really*?)

The **Native State**: the (awful) **Tertiary Structure**



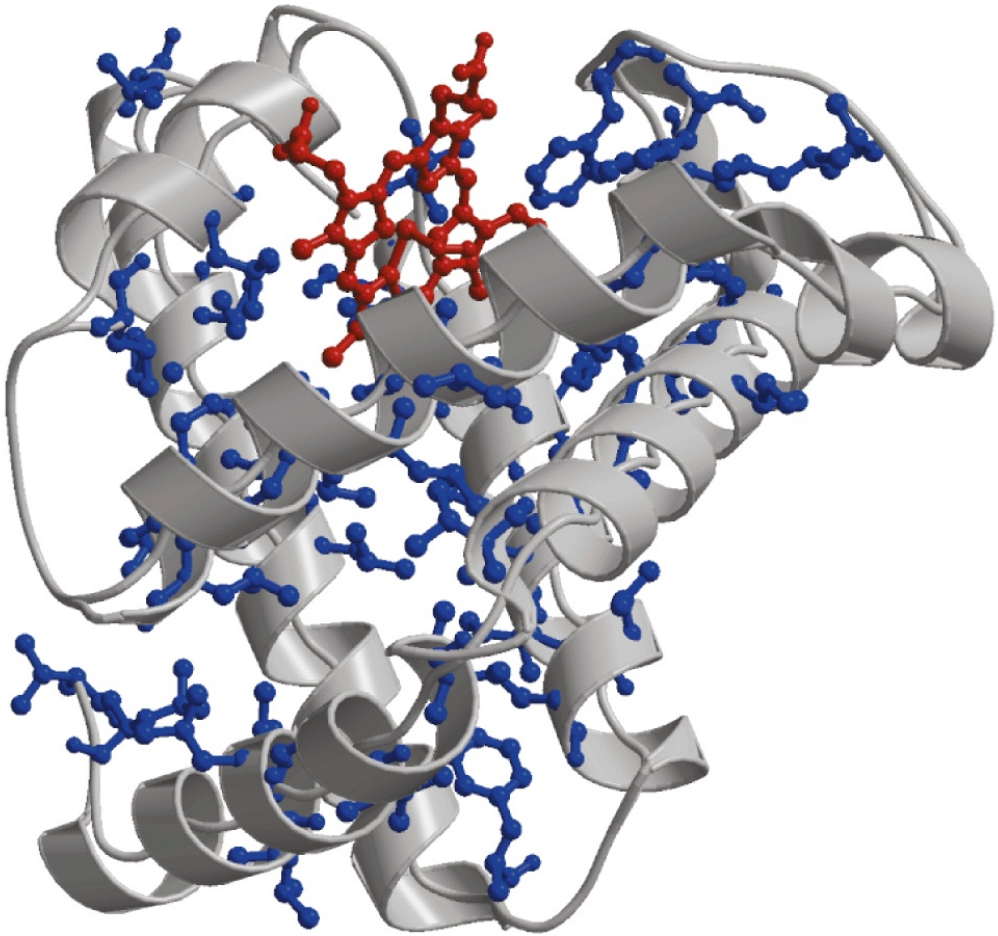
(c)

By the way: it lives at room temperature, so don't expect it to stay still! Look at the possible backbone configurations, as reconstructed through NMR...

(c)

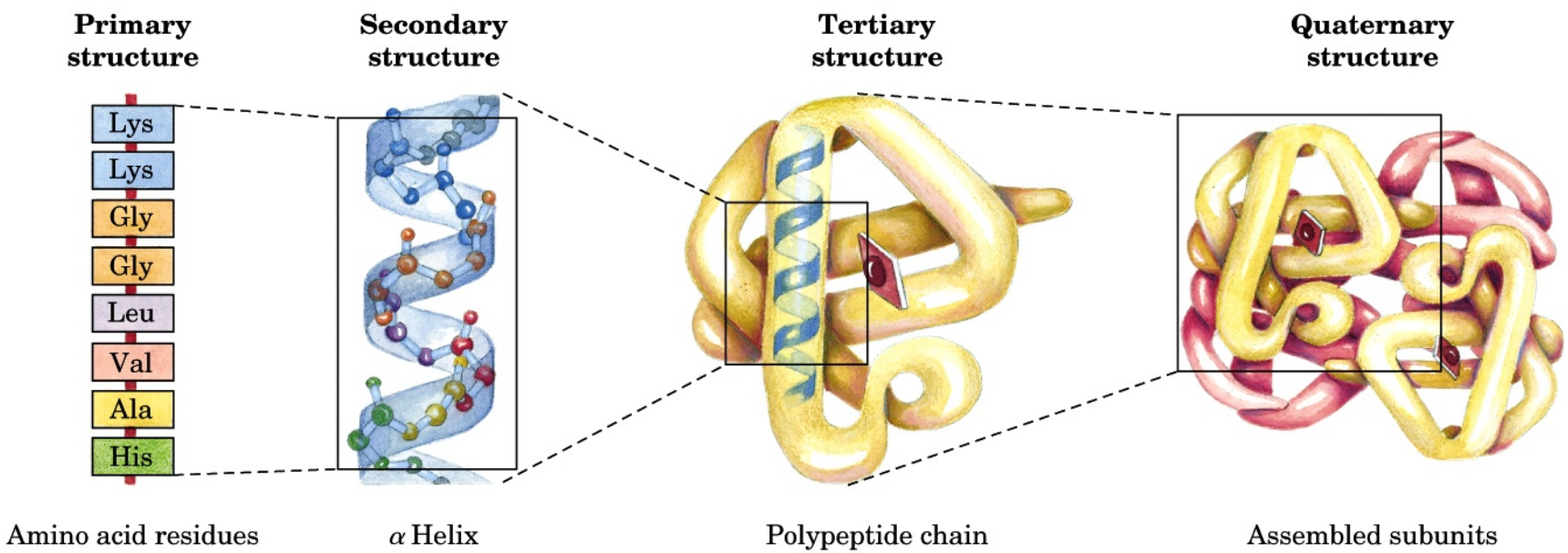
Yet it preserves the overall structure! Notice that a flexible structure is an advantage: it gives entropy (and hence, stability) and prompts for functional motion.

...and it contains some regularities! It has no symmetry, but it is made up assembling parts of roughly regular parts, with almost periodic backbone. Side chains protrude out of the backbone and fill densely the interior.



(d)

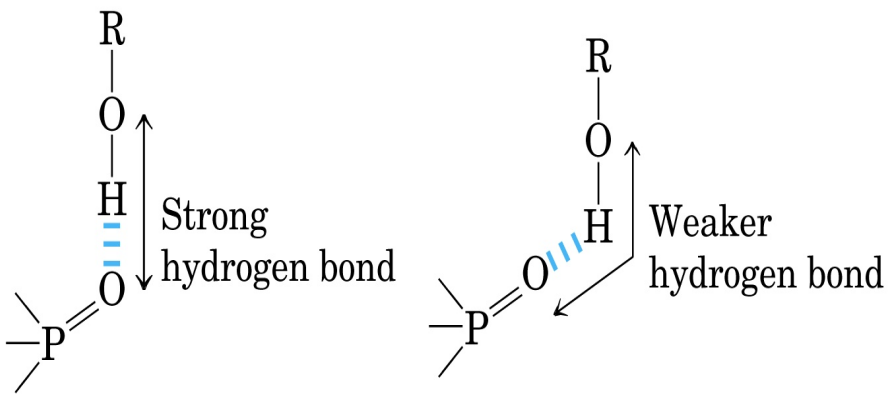
A closer look to the structure: let's see the details...

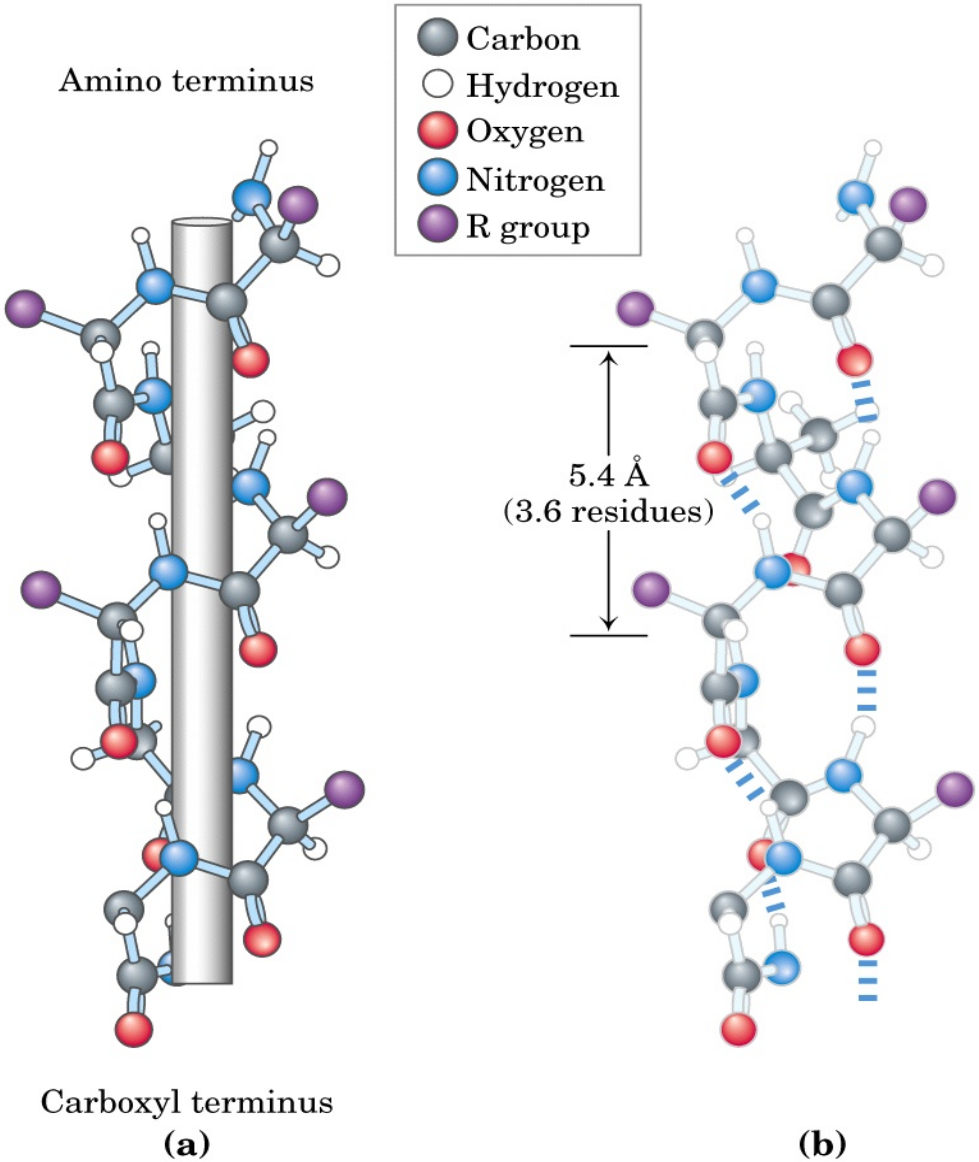


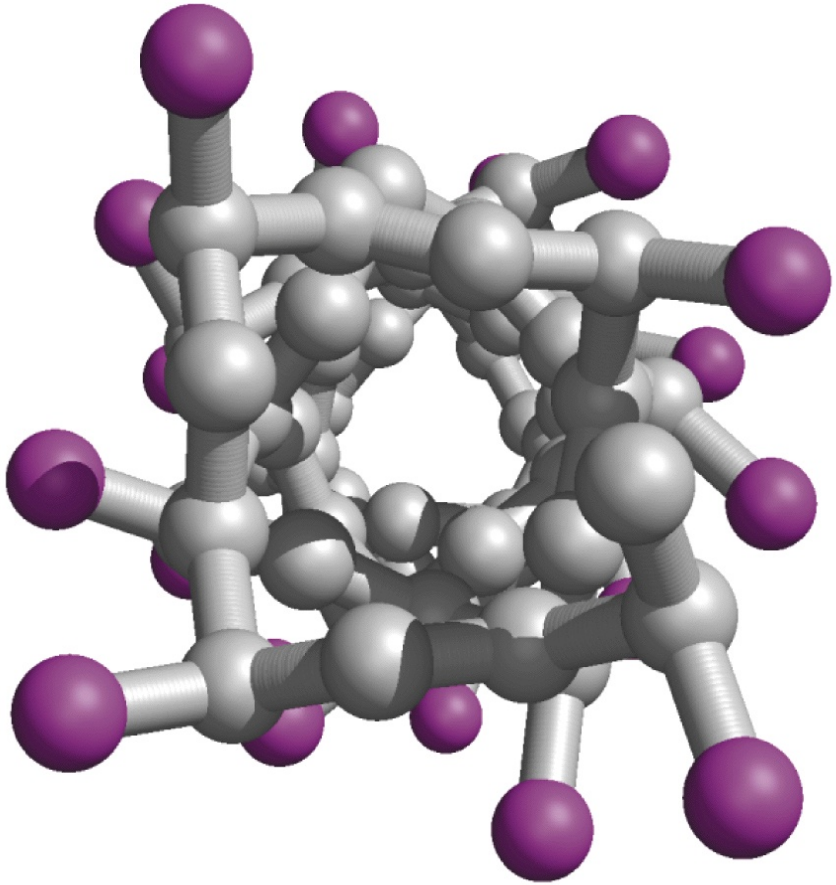
primary structure: the sequence of covalently bonded a.a.
secondary structure: hydrogen bonds
tertiary structure: 3-D overall arrangement of one chain
quaternary structure: assembly of several chains

α-helices and β-sheets are elements of regular Secondary Structure.

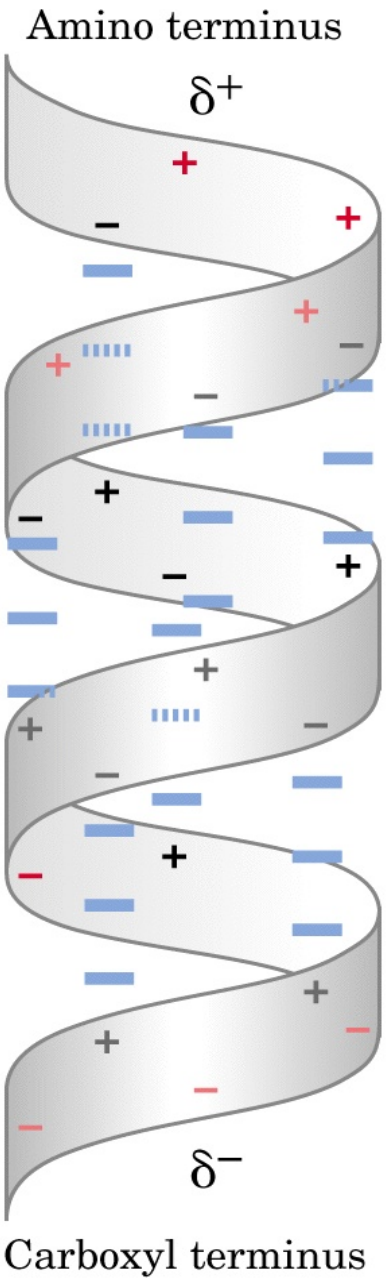
Regularity is provided by a network of highly directional hydrogen bonds.





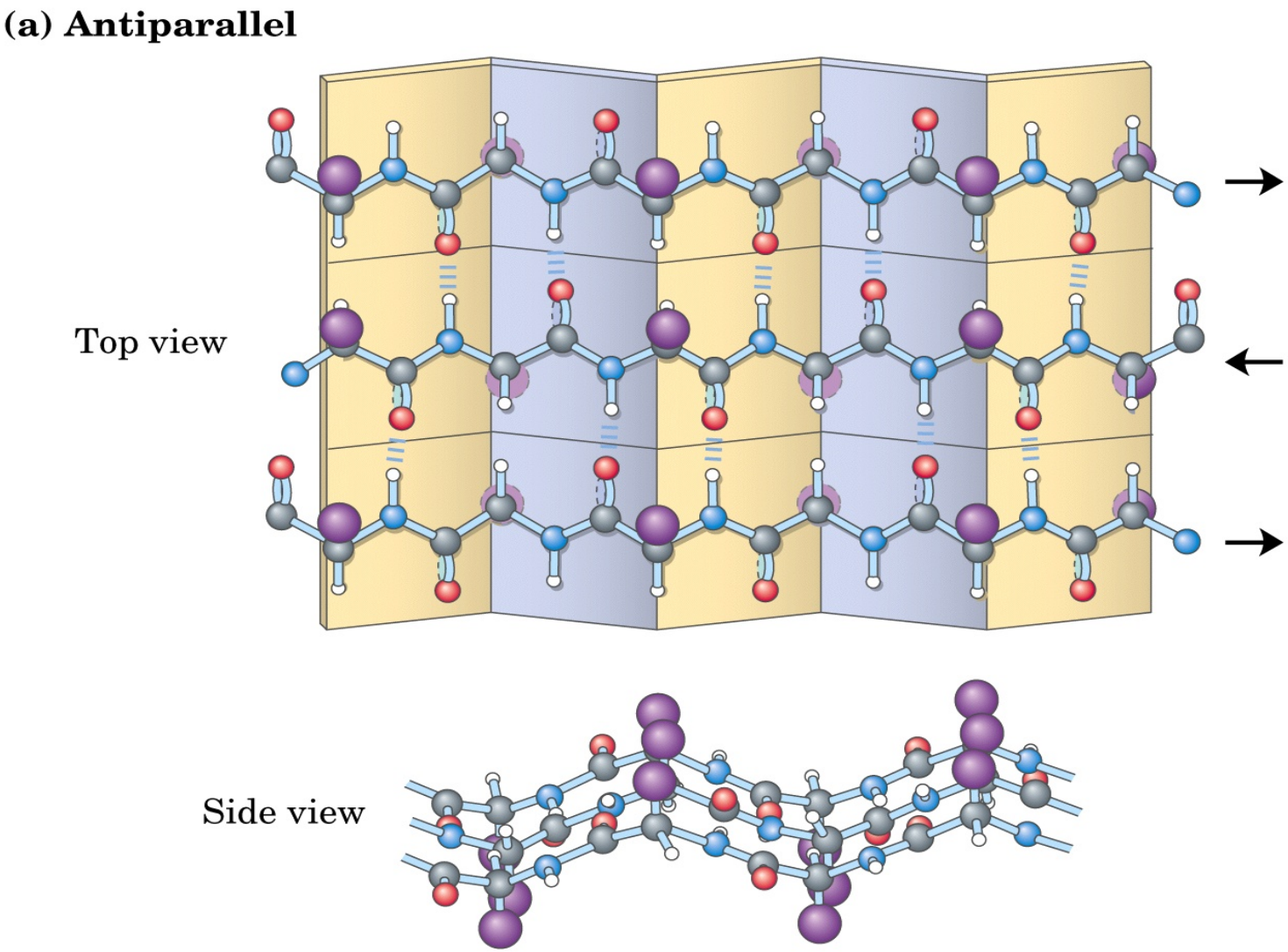


(c)



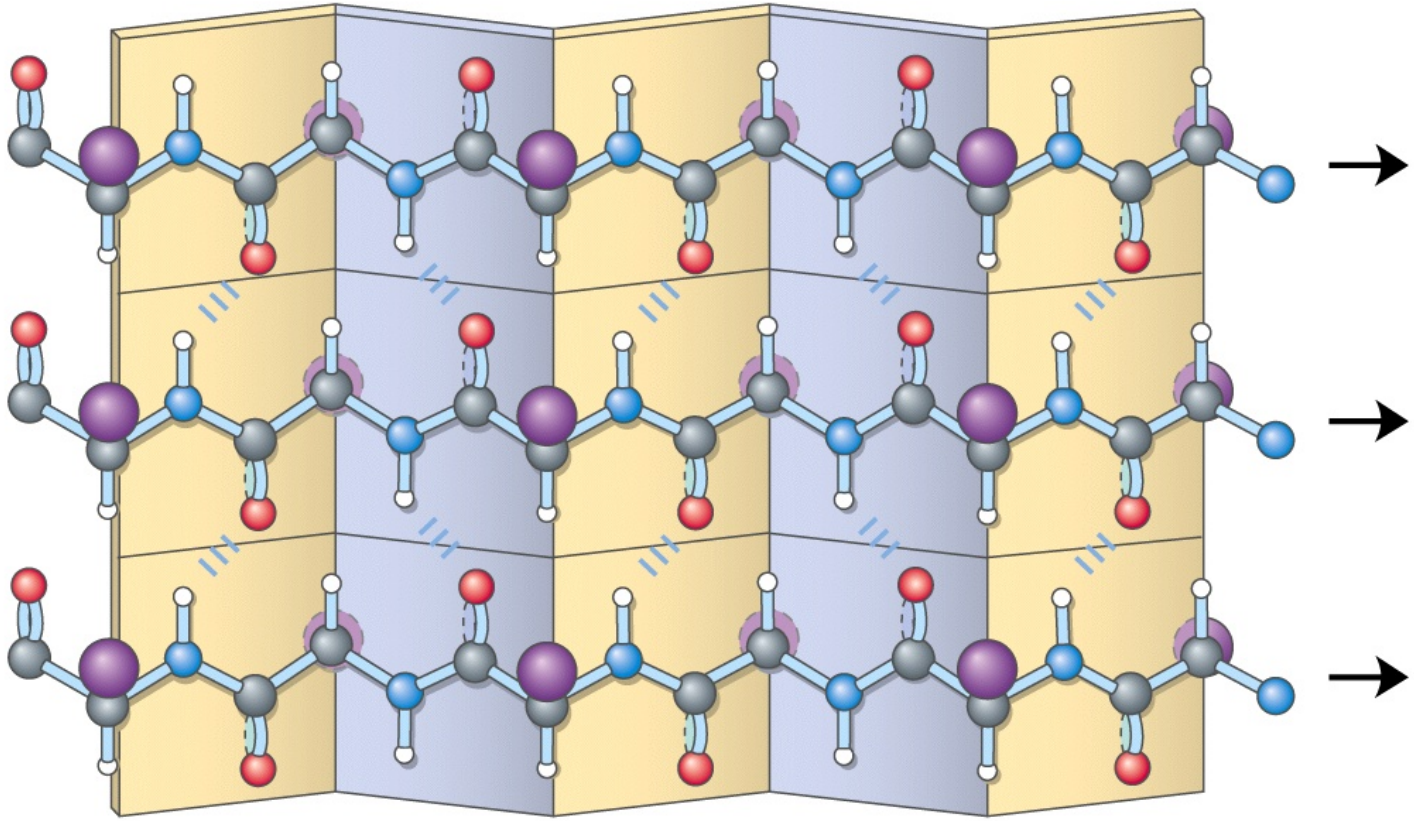
...helices have a dipolar moment

β-sheets are locally almost planar elements of structure. They can be parallel or antiparallel.

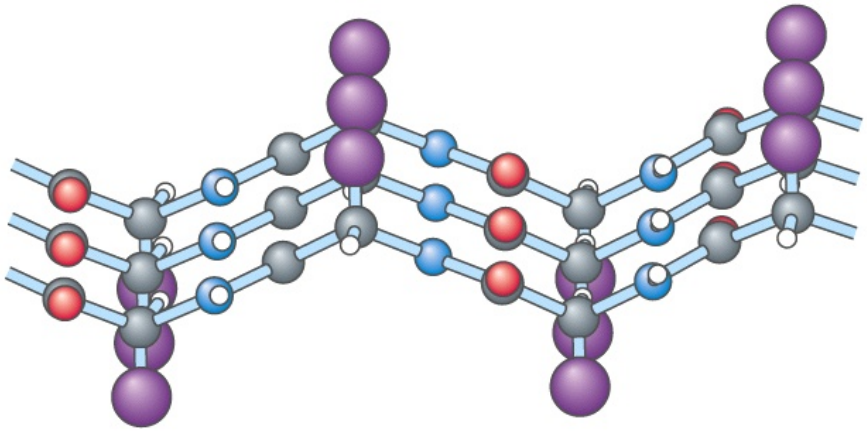


(b) Parallel

Top view



Side view



Some relevant questions:

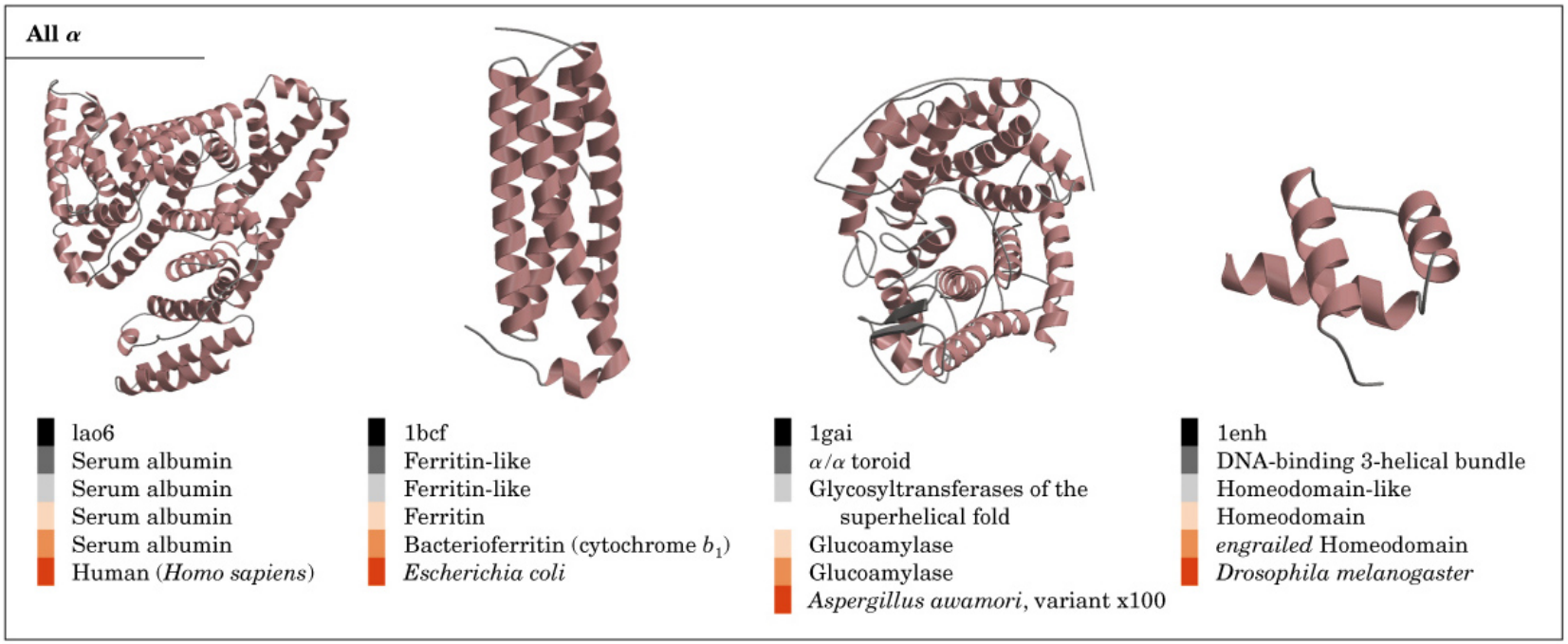
- How do secondary structure combine to produce tertiary structures?

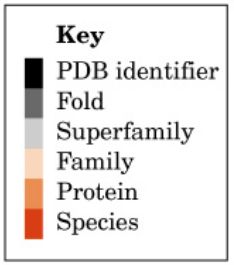
- Are tertiary structures arbitrary (in terms of any suitable geometrical indicator), or can only some structures exist?

- And if there are preferred structures, why? What are the rules? How can these rules be explained in thermodynamical and dynamical terms?

Let's see the experimental facts...

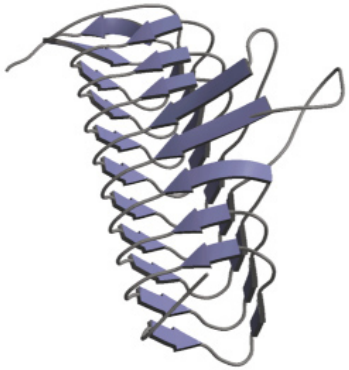
...It turns out that tertiary structures can be classified into classes and families:



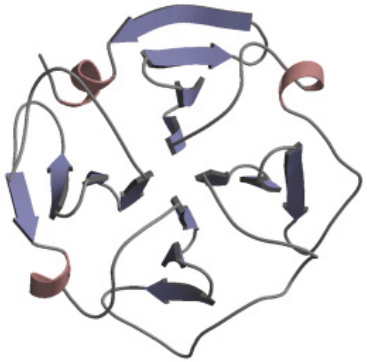


All β

1hoe
α-Amylase inhibitor
α-Amylase inhibitor
α-Amylase inhibitor
HOE-467A
Streptomyces tendae 4158

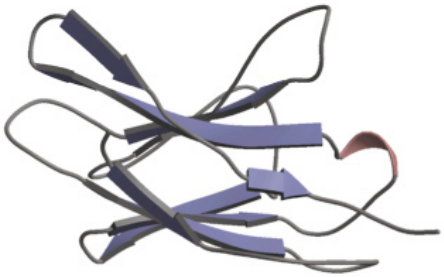


1lxa
Single-stranded left-handed β helix
Trimeric LpxA-like enzymes
UDP *N*-acetylglucosamine acyltransferase
UDP *N*-acetylglucosamine acyltransferase
Escherichia coli



1pex
Four-bladed β propeller
Hemopexin-like domain
Hemopexin-like domain
Collagenase-3 (MMP-13), carboxyl-terminal domain
Human (*Homo sapiens*)

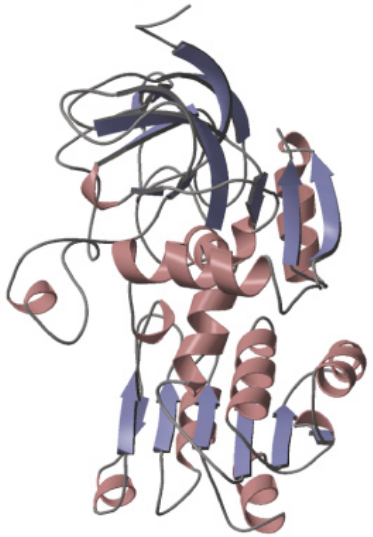
1jpc
β-Prism II
α-D-Mannose-specific plant lectins
α-D-Mannose-specific plant lectins
Lectin (agglutinin)
Snowdrop (*Galanthus nivalis*)



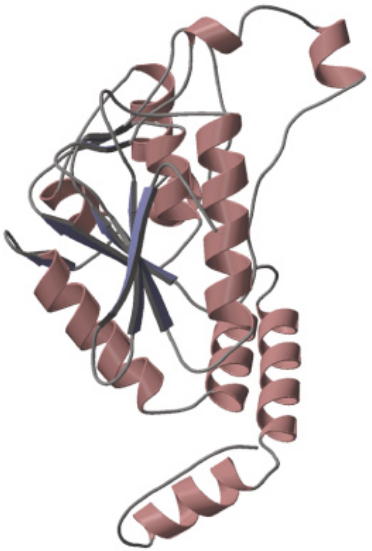
1cd8
Immunoglobulin-like β sandwich
Immunoglobulin
Antibody variable domain-like
CD8
Human (*Homo sapiens*)

Key
PDB identifier
Fold
Superfamily
Family
Protein
Species

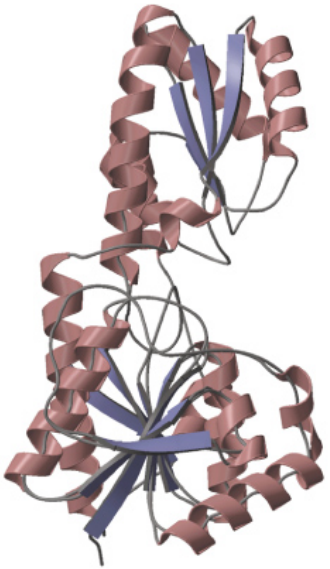
α/β



1deh
NAD(P)-binding Rossmann-fold domains
NAD(P)-binding Rossmann-fold domains
Alcohol/glucose dehydrogenases, carboxyl-terminal domain
Alcohol dehydrogenase
Human (*Homo sapiens*)



1dub
Crotonase-like
Crotonase-like
Crotonase-like
Enoyl-CoA hydratase
Rat (*Rattus norvegicus*)

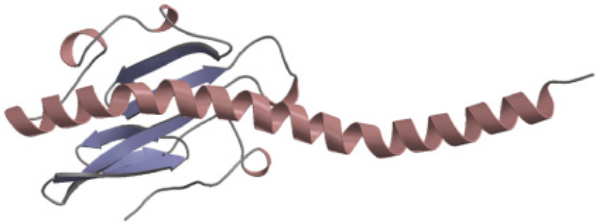


1pfk
Phosphofructokinase
Phosphofructokinase
Phosphofructokinase
Phosphofructokinase
Escherichia coli

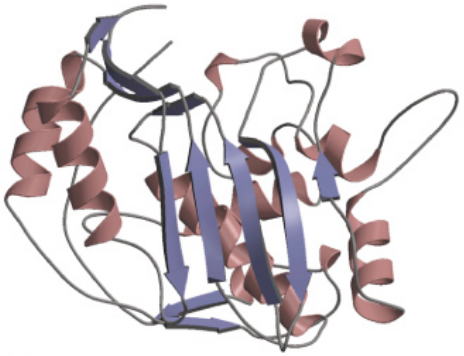
Key

PDB identifier
Fold
Superfamily
Family
Protein
Species

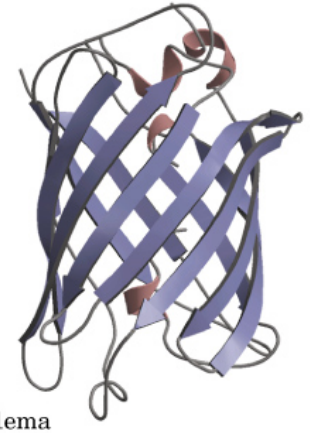
α + β



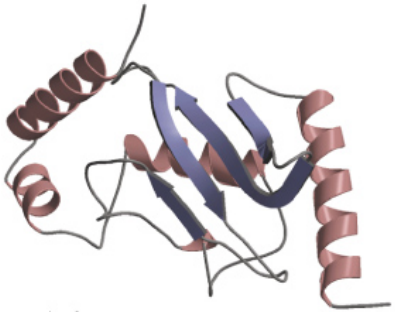
2pil
Pilin
Pilin
Pilin
Pilin
Neisseria gonorrhoeae



1syn
Thymidylate synthase
Thymidylate synthase
Thymidylate synthase
Thymidylate synthase
Escherichia coli



1ema
Green fluorescent protein
Green fluorescent protein
Green fluorescent protein
Green fluorescent protein
Jellyfish (*Aequorea victoria*)



1u9a
Ubiquitin-conjugating enzyme
Ubiquitin-conjugating enzyme
Ubibuitin-conjugating enzyme
Ubiquitin-conjugating enzyme
Human (*Homo sapiens*)

Key
PDB identifier
Fold
Superfamily
Family
Protein
Species

At present, millions of sequences:

CATH v3.1.0: Date 19-01-2007
N. of Domains: 93885 N. of Chains: 63453 N. of PDBs: 30028

C	A	T	H
Mainly Alpha	5	305	652
Mainly Beta	20	191	415
Alpha Beta	14	496	922
Few Sec.Str.	1	92	102
Total	40	1084	2091

CATH: hierarchical classification of protein domain structures, in Class(C), Architecture(A), Topology(T), Homologous superfamily (H).

A: overall orientation of secondary struct., independent of connectivities.
T: topological connections and numbers of secondary structures.
H: proteins with highly similar structures and functions.

- Observed proteins sequences are just a tiny fractions of all the possible sequences, and there is a wide degeneracy: many sequences fold on basically the same structure, and many structures share the same overall geometrical characteristics.

- Not all the folds are equally “populated”: some are the native structure of a wide number of sequences, and others of just a few.

Protein's native structure: an unlikely trade-off among

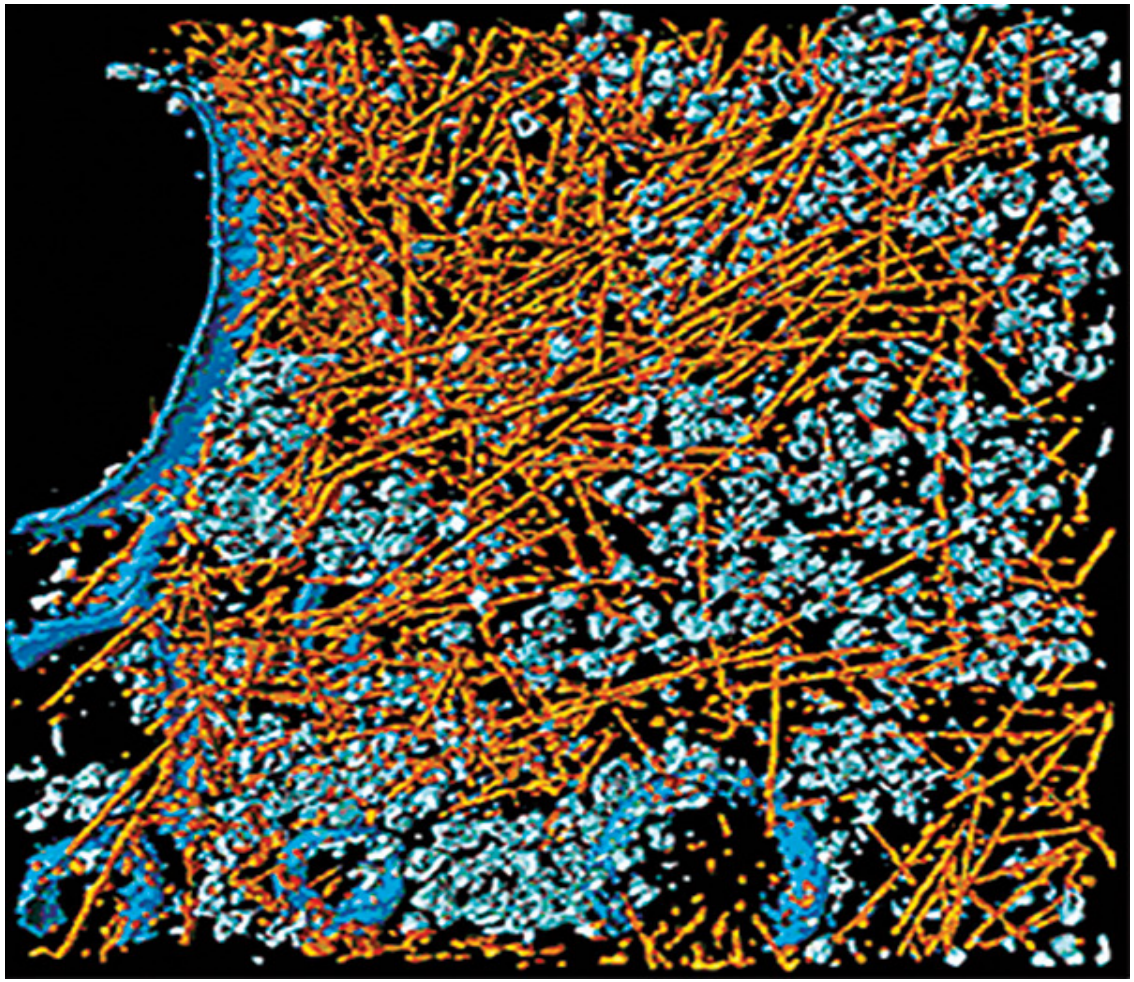
- stability

- function (constraints on stability, active site, allosterism,...)

- designability: a good structure, with a nice stability and capable of function, must be kinetically accessible in reasonable time.
  Why are some structures more populated? are they more designable?

- evolution: dynamics in sequence space is not ergodic, and we don't even know if it is ergodic in a subspace.
  Is evolutionary dynamics an optimization dynamics?

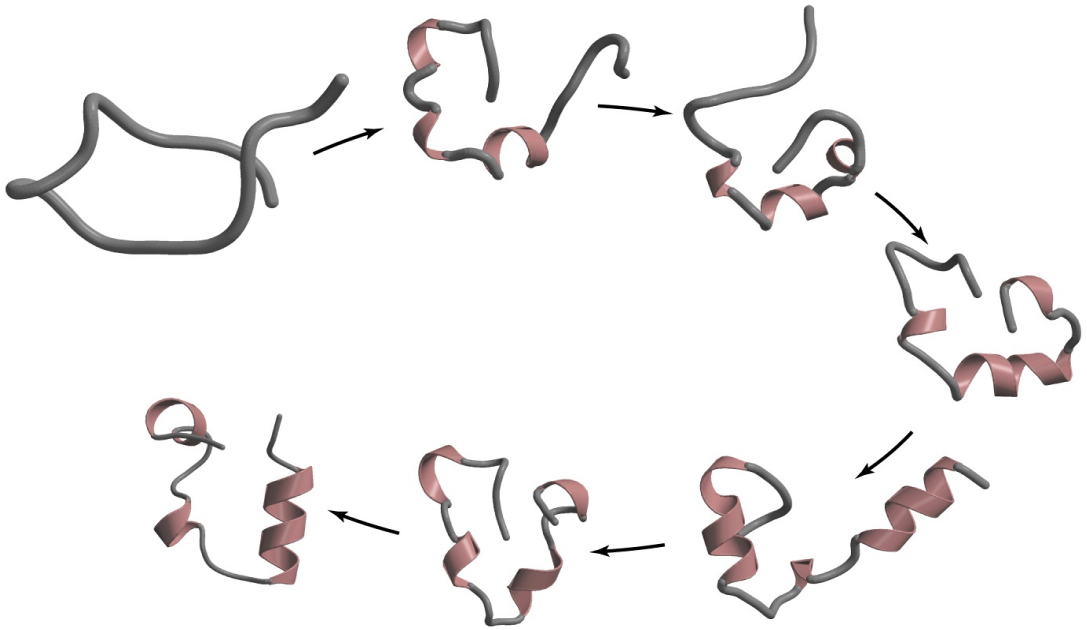
Let's go back to Protein folding:
in vivo: molecular crowding. Here is how a cell looks like...



“Chaperons” are often needed...

But in vitro: it is reproducible (at least for proteins<300 res).

Take a protein, “denaturate” it (by heating, adding acids, adding denaturants...): biological activity disappears.
But: restore physiological condition, and the protein will be active again.



Anfinsen's discovery:
1) Proteins “work” just when in their native state.
2) The native structure and the whole folding process is encoded in the sequence of aminoacids, and in nothing else.

Folding Problem(s):

1) native structure prediction from sequence

2) prediction of the folding process from denatured state to native: characterization of the unfolded state, intermediate states (are there?), folding rates,...

3) sequence design ("inverse folding" problem): given a structure, find the sequence folding on it

(4) protein-protein interactions during folding: interactions with chaperons, aggregation, misfolding, etc.)

Folding as a single molecule process:

Time scales:

10^{-14} -10^{-13} s: vibration of of a covalent bond

10^{-11} -10^{-10} s: backbone movement

10^{-7} -10^{-6} s: alpha helices

10^{-6} -10^{-4} s: beta sheets

10^{-5} s - 1 s: folding time

Stability: native state is only marginally stable with respect to denatured: 10 Kcal/mol, that is $\approx$ 20 $k_B T$. This means that if each residue were independent from the others, its stability would be only 20 $k_B T/N \approx 0.1\ k_B T$: also in native conditions, it would pass a lot of time unfolded. Cooperative behavior of the residues: folding is a sort of first-order transition, with a jump in enthalpy between the native and the denatured state.

Folding as a single molecule process: let's look at **kinetics**

Anfinsen's discovery of an essentially thermodynamic nature of protein folding is nice, but: Consider a single protein, in a unfolded conformation and suddenly restore refolding environment: it will reach the native state.

But how?

Proteins are flexible, monomers are small, and in solution: these changes will be ruled by brownian motion.

Levinthal paradox:
consider a protein of N residues: if each peptide unit has k conformation, the native state will be one of the k^N conformations of the protein.
Taking $k \approx 8$ and $N \approx 100$, for a random search with moves every 10^{-13} s,
folding time >> age of the Universe

Solution of this “paradox”:

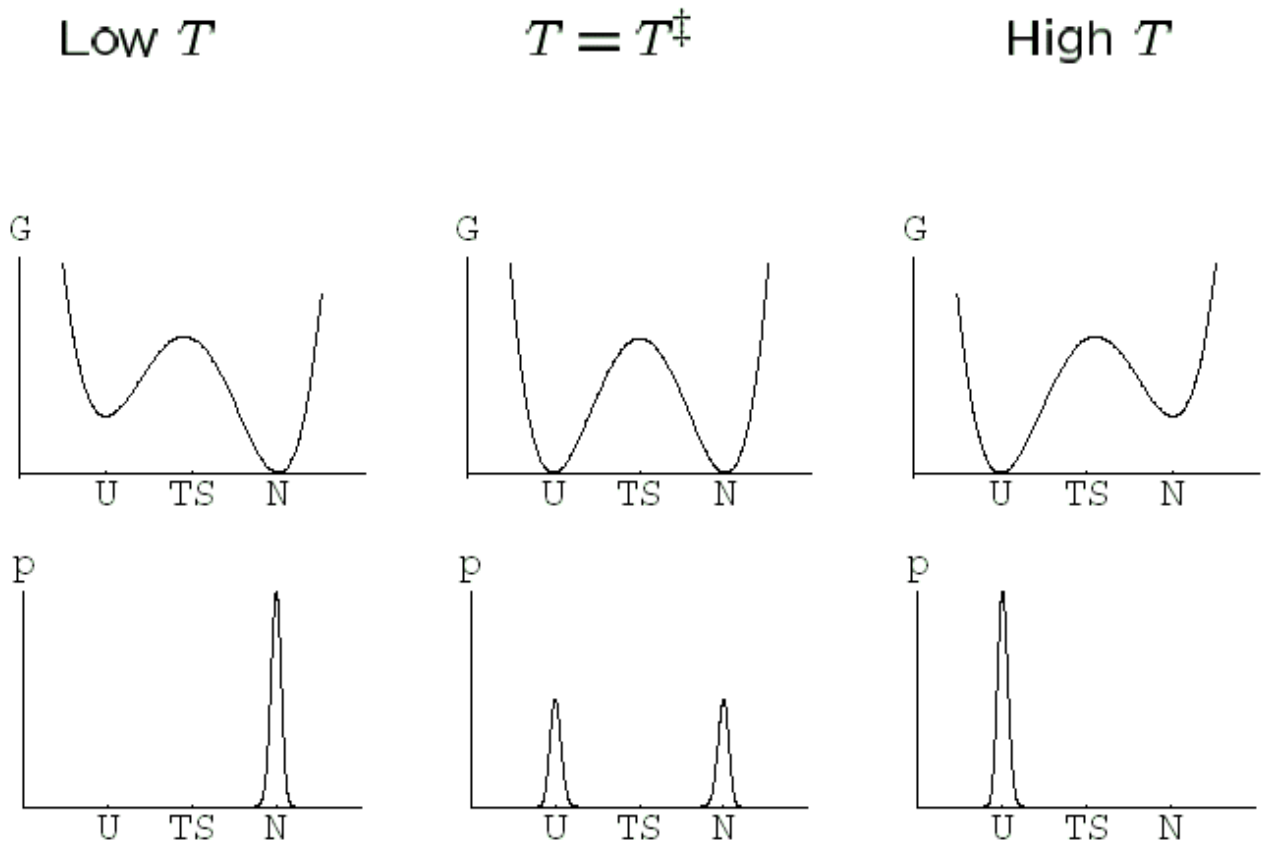
folding is not like a random search of the holes in a golf field:
there must be a “driving force” reducing the dynamical exploration
to a reasonable number of microstates.

If we think about the potential energy landscape of the protein,
we realize that minima and saddles cannot be arbitrary.

Indeed, very often experiments reveal a two-state behavior in kinetics:
there is no transient phase between depopulation of the unfolded state
and population of the native one.

Experimentally: few populated states in Thermodynamics an Kinetics

Two-state thermodynamics:



$$I_{obs}(T) = \frac{I_N(T)e^{-\beta G_N} + I_U(T)e^{-\beta G_U}}{e^{-\beta G_N} + e^{-\beta G_U}} = I_N(T)f_N(T) + I_U(T)f_U(T)$$

$f_{N,U}$=fraction native, unfolded

Experimentally: few states Kinetics. Let's see **Two-State:**

Consider the process $U \underset{k_u}{\overset{k_f}{\rightleftharpoons}} N$

In general,

$$\frac{df_N}{dt} = -k_u(t) f_N(t) + k_f(t) f_u(t)$$

If $k_{f,u}$ constants: two-state kinetics:

$$\frac{f_N(t) - f_{N,\infty}}{f_{N,0} - f_{N,\infty}} = e^{-kt}$$

with $k = k_f + k_u$, $f_{N,0} = f_N(t=0)$, etc

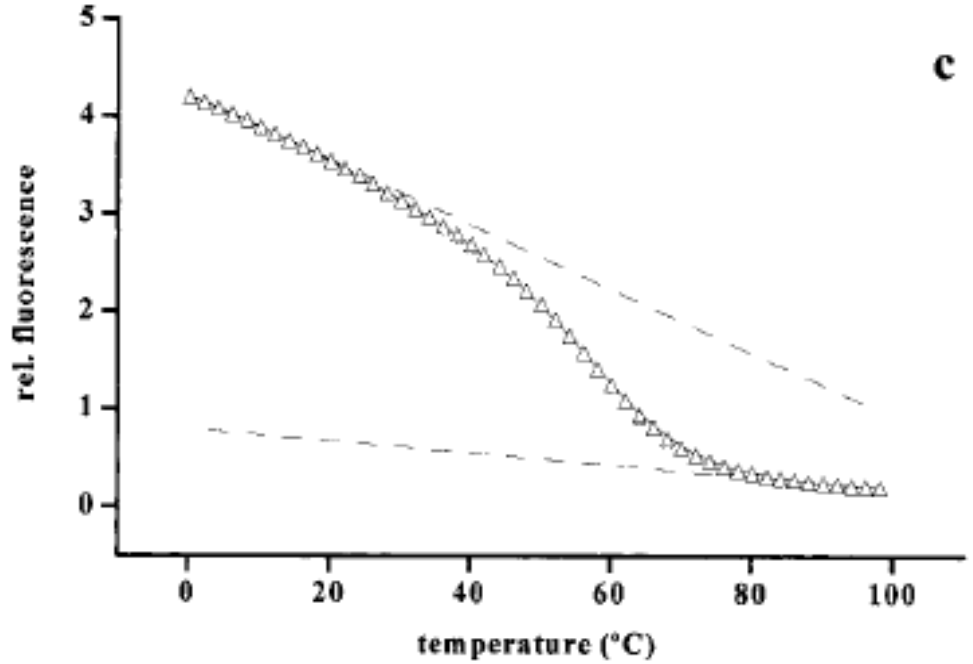
Notice that the few-state behavior is so common that in general biologists postulate them to interpret their data!!!

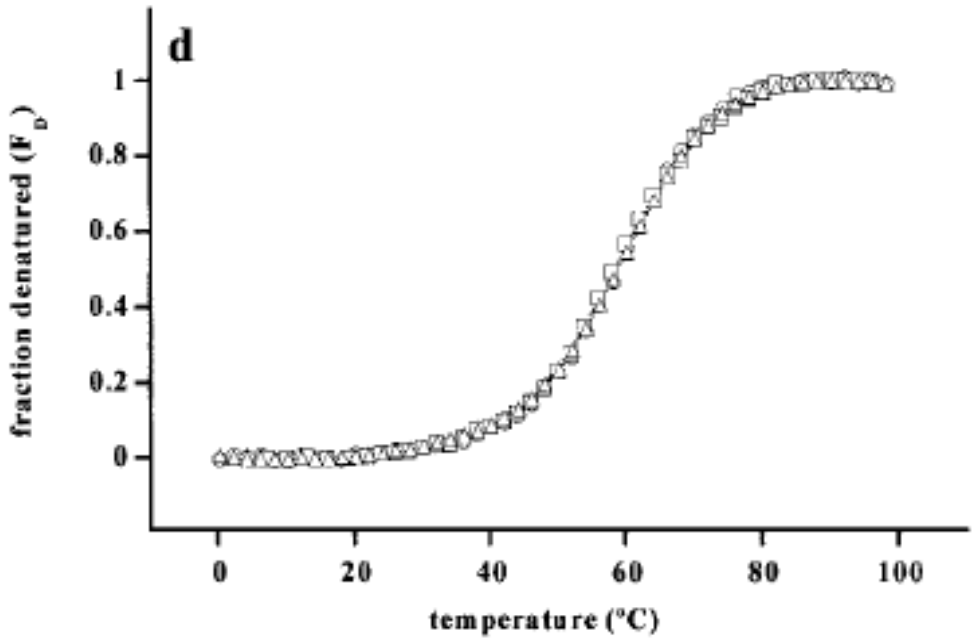
In thermodynamics (see Adrian Velazquez's lectures):

assuming that just N,U states exist, and that their "baseline" is $I_{N,U}(T)$, find K to fit the experimental signal:

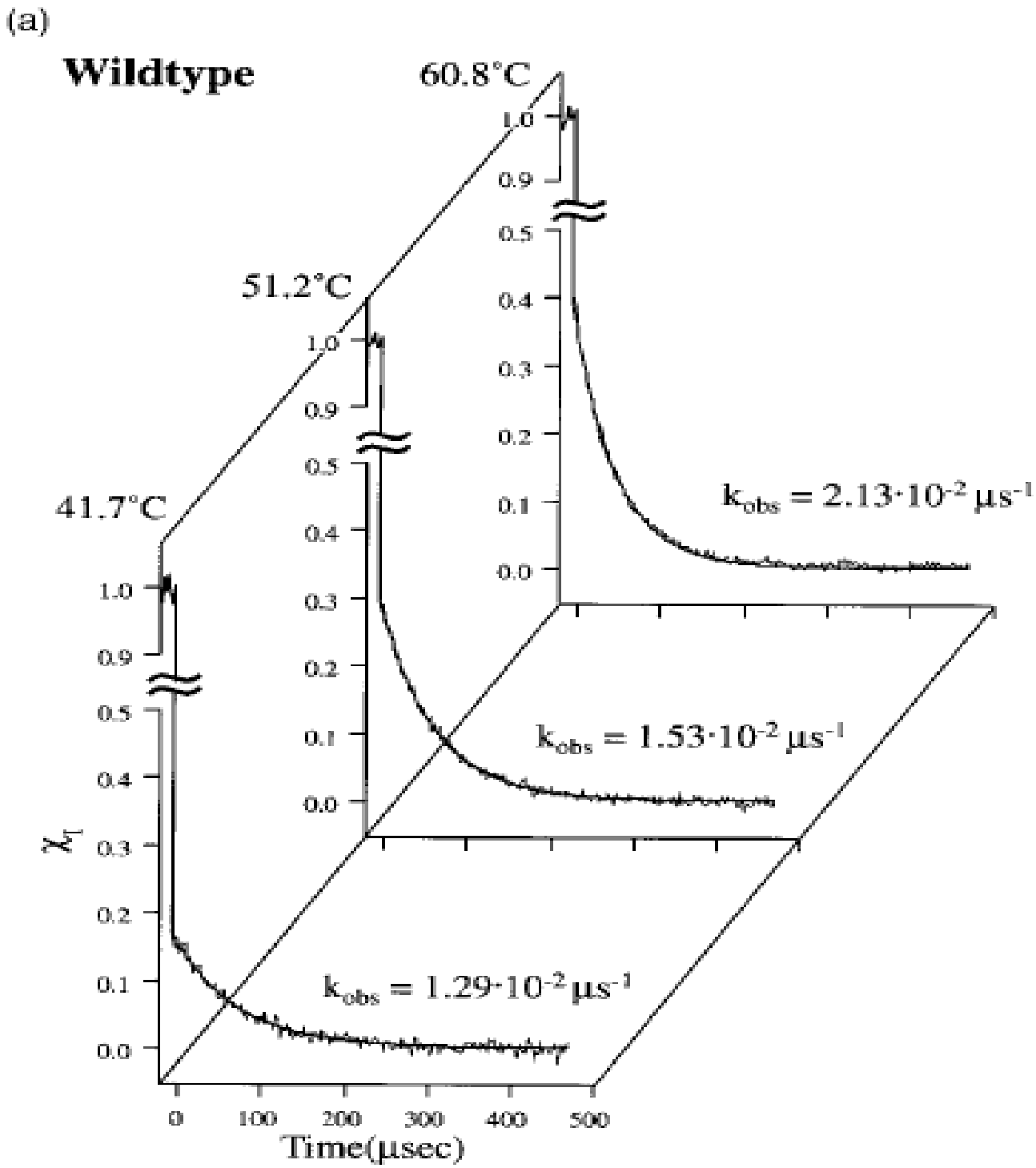
$$I_{\exp}(T)=\frac{I_N(T)e^{-\beta G_N}+I_U(T)e^{-\beta G_U}}{e^{-\beta G_N}+e^{-\beta G_U}}=\frac{I_N(T)K+I_U(T)}{K+1}$$

where $K(T)=e^{-\beta(G_N-G_U)}=\frac{f_N}{f_U}=\frac{k_f}{k_u}$


c
rel. fluorescence
5
4
3
2
1
0
0
20
40
60
80
100
temperature (°C)


d
fraction denatured (F_D)
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
temperature (°C)

In kinetics,


(a)
Wildtype
60.8°C
51.2°C
41.7°C
$k_{obs} = 2.13 \cdot 10^{-2}\ \mu s^{-1}$
$k_{obs} = 1.53 \cdot 10^{-2}\ \mu s^{-1}$
$k_{obs} = 1.29 \cdot 10^{-2}\ \mu s^{-1}$
χ_T
Time(μsec)

In kinetics, they fit the signal to a single exponential (and in case of poor quality, to a two-three exponential, or to a stretched exponential). For a two-state system, we have:

$$I_{\exp}(t) = \left(I_N f_{N,\infty} + I_U f_{U,\infty}\right) + \left[I_N (f_{N,0} - f_{N,\infty}) + I_U (f_{U,0} - f_{U,\infty})\right] e^{-kt}$$

Are these simple behaviors easily explained in terms of the interactions involved in the folding process? Let's see...

Forces in protein folding

The minimal ingredients for compactification to a unique structure:

- self avoidance (obvious: but also mandatory for uniqueness)
- attractive effective monomer-monomer interactions (so to allow the low temperature collapse).
- ...and perhaps more.

Do we have them?

Let's have a look at the interactions in the game:

covalent bonds:	50-150 Kcal/mol ($\gg k_B T$: primary seq. is fixed)
hydrogen bonds:	5-8 Kcal/mol (non-polar solvent)
	1-2 Kcal/mol (polar solvent)
Van der Waals:	1 Kcal/mol
Coulomb:	1 Kcal/mol

These are true, direct, interactions, especially effective at low distances ($\approx$0.1 nm). However, it 's common opinion that hydrophobicity is the major leading force for folding.

Hydrophobicity: basic facts:

Protein folding take place in water: to establish stability of Native vs Unfolded, not only energy but also entropy must be considered !

$$\Delta F_{N-U} = \Delta H_{prot} + \Delta H_{solv} - T \Delta S_{prot} - T \Delta S_{solv}$$

$\Delta H_{prot} > 0$ Van der Waals in protein interior are weaker than interaction with water, even of apolar species

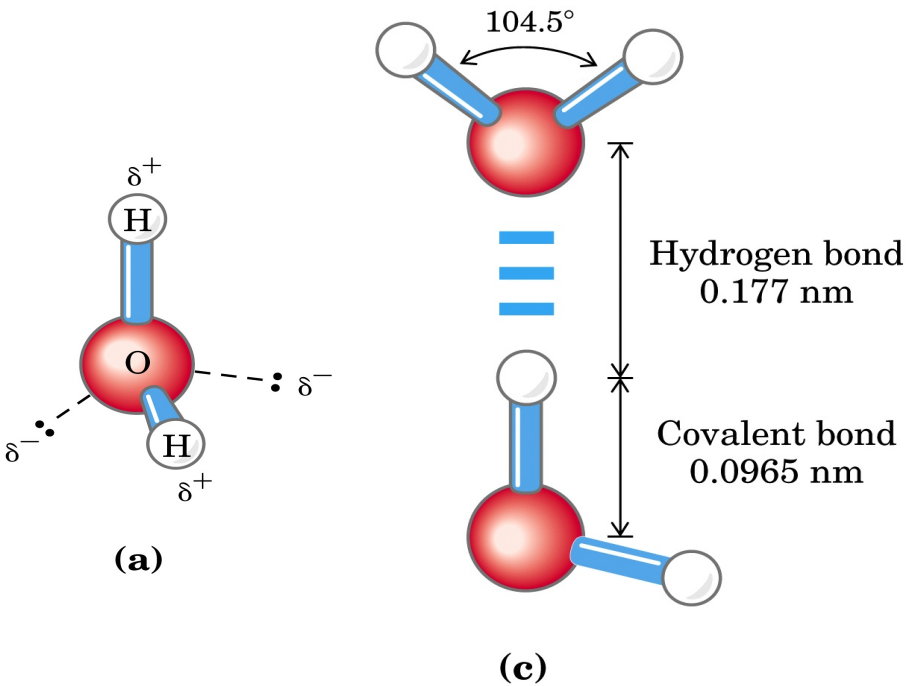
$\Delta H_{solv} < 0$ Water makes more H-bonds with itself if protein is folded

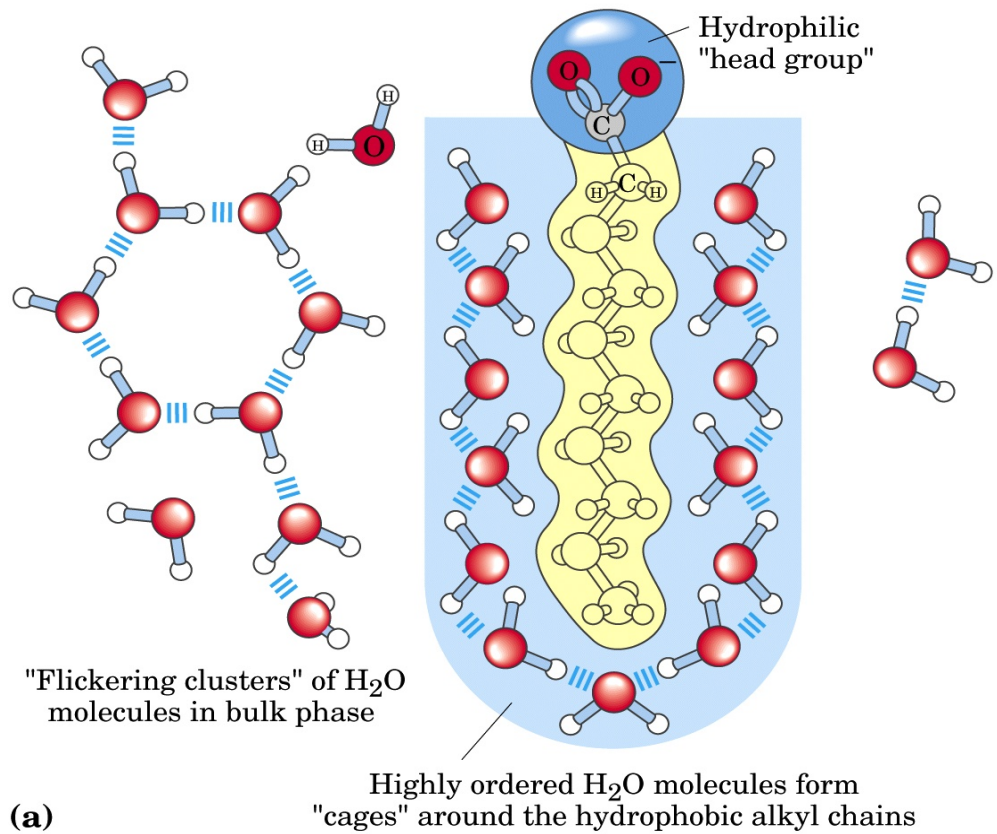
$\Delta S_{prot} < 0$ Obviously, a native compact chain has much smaller entropy than an swollen one

$\Delta S_{solv} > 0$ To keep its network of Hydrogen bonds, water gets ordered around a nonpolar specie: low water entropy in unfolded state!

Hydrophobic interactions:

nonpolar molecules induce ordering in water to preserve hydrogen bonds:
entropy favors effective attraction between nonpolar (hydrophobic) residues

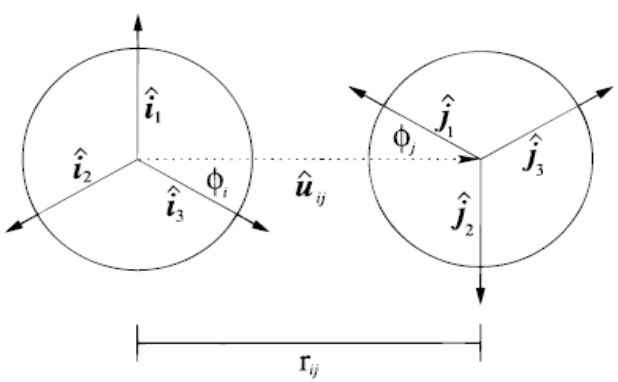


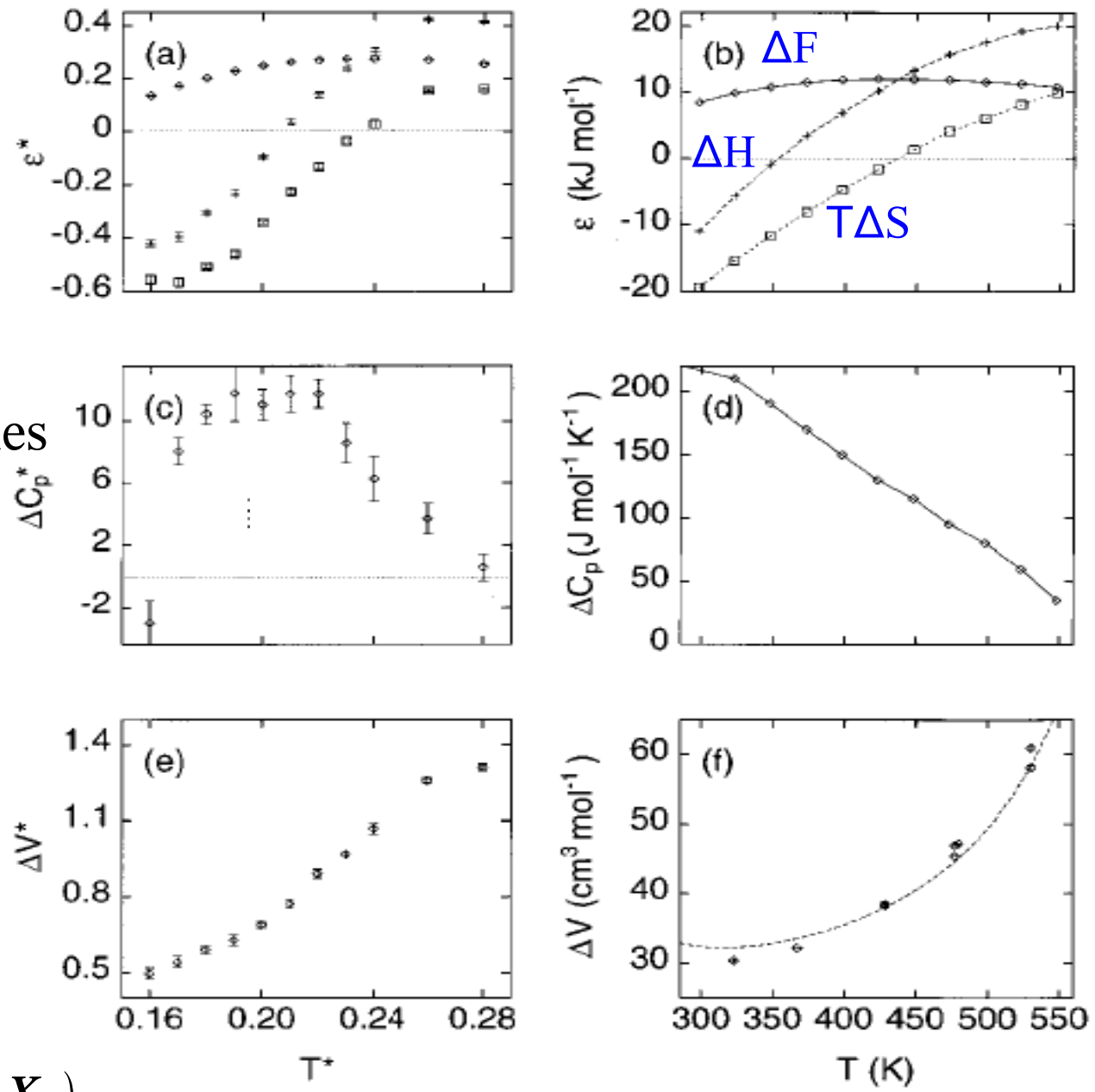


Silverstein et al, 3172 J. Am. Chem. Soc., Vol. 120, No. 13, 1998

Experiment:
Transfer of a nonpolar specie (Argon) from gas into water

Model:
argon: circle, Lennard-Jones
water: LJ+HB





$$U(\boldsymbol{X}_i, \boldsymbol{X}_j) = U_{LJ}(r_{ij}) + U_{HB}(\boldsymbol{X}_i, \boldsymbol{X}_j)$$

Experimental

What we have learnt so far:

Thermodynamics (equilibrium)

- folding is substantially an equilibrium thermodynamic process: an external parameter, (T, [Denaturant], pH,...) acting as an intensive variable, shifts the thermodynamic equilibrium towards one of two, or some, states.

- one of these states, the Native State, is peculiar for being:
  * functional
  * a collection of "few" microstates, close in geometry
  * geometrically structured, with "regular" secondary structures and limited assembling preferences.

  The others phases are:
  * an unfolded state, whose radius of gyration depends on the conditions (it may be swollen or keep residual local structure)
  * possibly some equilibrium intermediates, with low population

What we have learnt so far (cont'd):

Thermodynamics (equilibrium)

- These equilibrium phases are separated by first-order-like transitions: cooperativity plays a fundamental role: a barrier is present separating N from U. Folding implies jumping over this barrier.

- Several interactions involved. Entropic “interactions” are relevant (hydrophobicity)

Kinetics

- Microscopic landscape must be “tuned” for easy and robust dynamics

Born to Function:

Native state must be:

- global thermodynamical minimum (or a long lived metastable: a few cases exist)

- quickly accessible by dynamics in a reasonable time scale (ms ÷s)

- stable against denaturation, but not too much (function implies switching between states, degradation must be possible)

- "resistant" to mutations: highly designable!
  (A polymer of 100 a.a can have 10^{130} sequences: if structures weren't highly designable, life wouldn't have evolved)

What would we like to predict with a model:

- stability of the native state

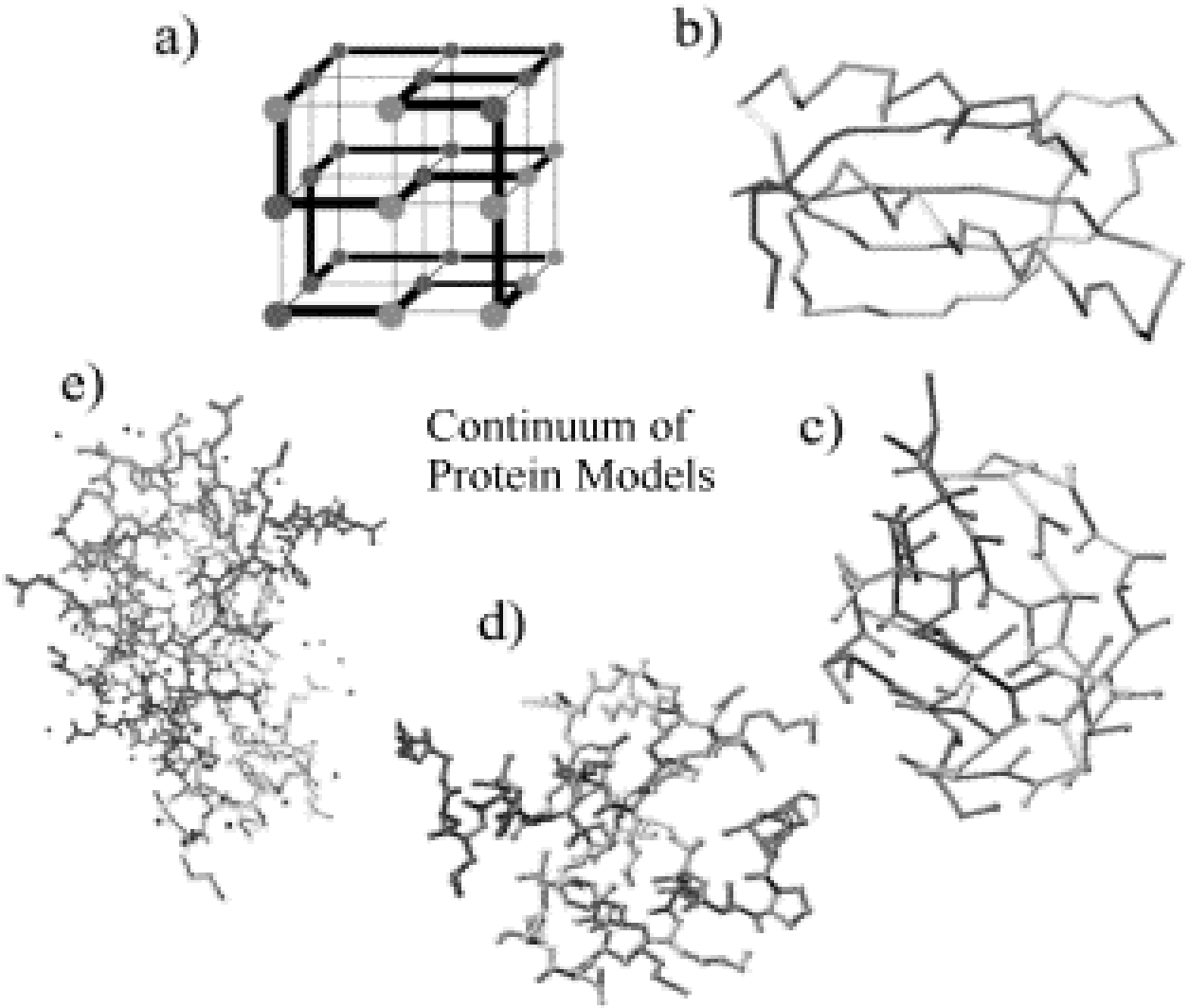
- characterization of the important steps in the kinetics: what are the “bottlenecks”?

Difficulties:

- the first requires the evaluation of free-energies, that is: a global knowledge of the energy landscape

- the second requires the characterization of the “saddle points” and of unlikely conformations, very difficult to sample

Theoretical models: from lattice to all-atoms description.



How can we model all this?

Relevant numbers: $N_{ato,\ prot} >\approx 1000$, $N_{ato,\ solv} \approx 10\ N_{ato,\ prot}$: tens of thousands of atoms to simulate.

- In principle, the most microscopic events (e.g formation of H-bonds), would need a quantum mechanical approach. But, even with approximations of Density Functional Theory, etc this is computationally out of reach.
  (at present, QM limited to a subset of atoms, in a QM/MD scheme)

- Simplified all-atoms description: atoms as classic spheres, classical Newton equations to move them (Molecular Dynamics)

Further possible simplification:
divide variables into slow and fast, integrate fast degrees of freedom.

From the (potential energy) part of the partition function:

$$Z_V = \int d\boldsymbol{x}\, d\boldsymbol{X} \exp(-\beta V(\boldsymbol{x}, \boldsymbol{X}))$$

For each value of the slow **X**, one can assume equilibrium for the fast variables **x** , to obtain an effective potential for the slow variables:

$$V_{eff}(\boldsymbol{X}) = -k_b T \ln\left(\int d\boldsymbol{x} \exp(-\beta V(\boldsymbol{x}, \boldsymbol{X}))\right)$$

This is very common to “remove” water and work with just the protein.

All-atoms empirical potentials (CHARMM, AMBER,...)

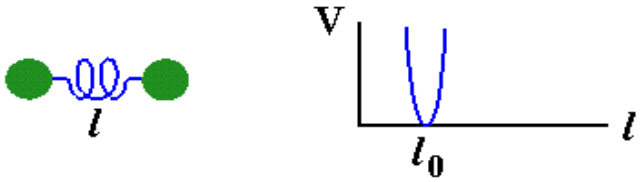
Empirical Potential Energy Function

$$V(\vec{R}) = V_{bonded}(\vec{R}) + V_{nonbonded}(\vec{R})$$

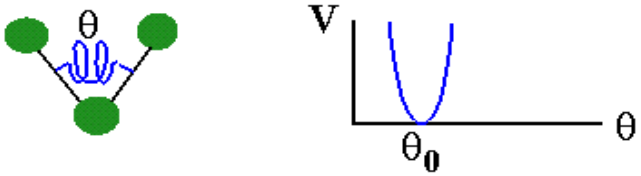
$$V_{bonded}(\vec{R}) = \sum_{bonds} k_l(l - l_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{impropers} k_\omega(\omega - \omega_0)^2 + \sum_{torsions} A_n\left[1 + cos(n\phi - \phi_0)\right]$$

$$V_{nonbonded}(\vec{R}) = \sum_{i<j} \left(\varepsilon_{ij}\left[\left(\frac{r_{ij}^{min}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{min}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_r\epsilon_0 r_{ij}}\right)$$

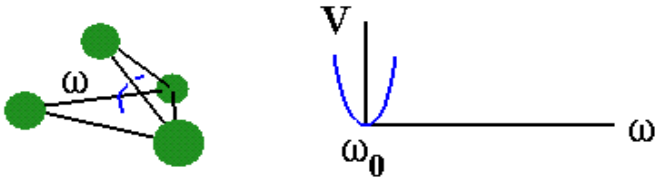
Bonds



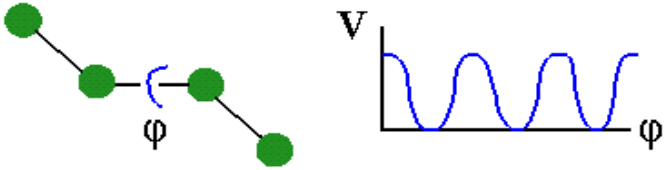
Angles



Improper Dihedrals

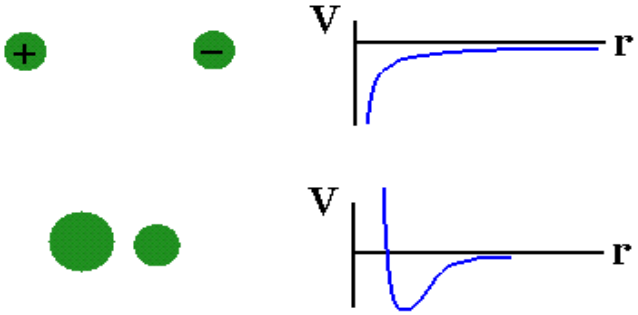


Torsions



Electrostatics

van der Waals



Applications:

- When a quantitative prediction is required, they are the best one can use.

- Very useful for structure refinement (adding constraints from experiments) and to simulate high T unfolding

- adding ad-hoc potentials, with extra statistical terms, they are useful for “ab-initio” structure prediction

Remarks:
- they are extremely “popular”: everyone tend to believe them.
- they represent the approach with more chances to improve

The above reasons explain projects like IBM's BlueGene, or Riken MDGrape (petaflop dedicated computer)

Results: realistic, but short

- MD simulations with "realistic" force fields can describe a ns life of a small protein (45 residue), at physiological T with explicit solvent, 10÷100 ns with implicit solvent.
  (NB: bond formation: ~ 0.1 ps, so that time step ~10^{-15} s is necessary. Folding times > 1ms approx)

- just a few runs: little statistics. Equilibrium properties?

Remarks:
simulations are reliable, if potential is reliable! Though realistic, these potential are empirical. Even in the all-atoms, explicit solvent case, there is a number of implicit assumptions:
- two-body terms
- transferability of parameters calculated for one molecule to another.

The description of an atom involves different parameters according to the environment (different hybridization, different chemical bonds,...)

Coarse-grained model and "inferred potentials": resort to

-Simpler description of chain: Cα, Cβ, UNRES, ...

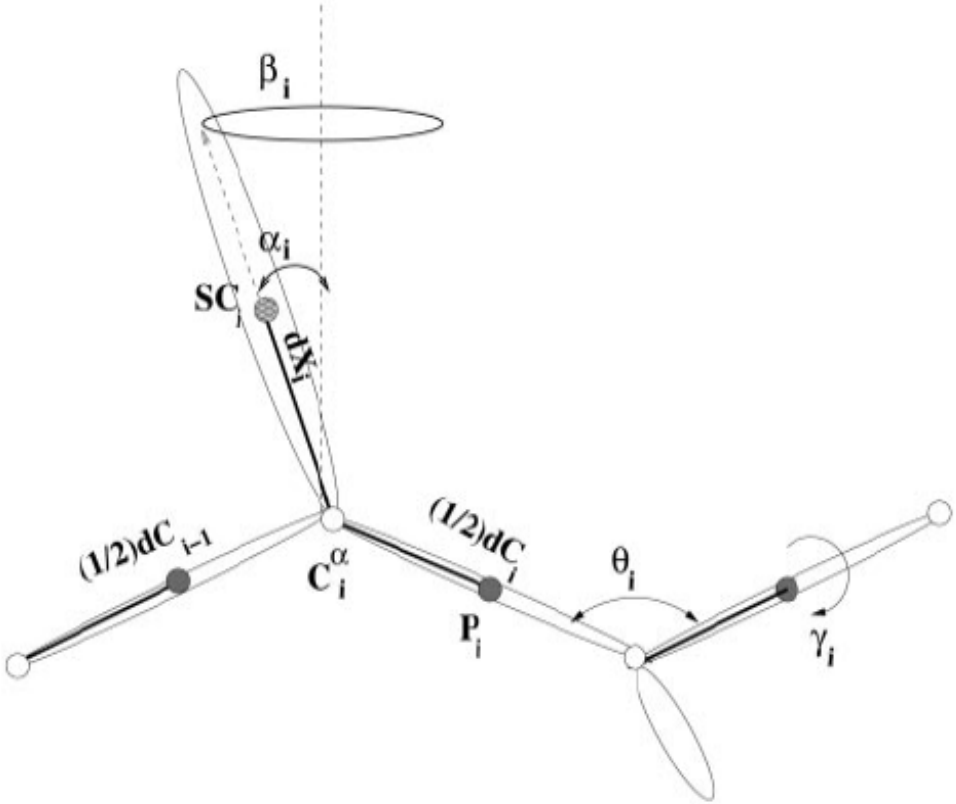
-Potentials inferred resorting to a database of structures:

Potential of mean force: derived from frequency in a library of conformations: from the frequency of features(contacts, conformations, etc), infer a potential:
for instance: $\varepsilon(A,B)$ between aa. of type A and B can be inferred from the frequency f(A,B) by which they appear in "contact" in the database
$\varepsilon(A,B) = -\text{ const } \ln(f(A,B))$

Learning by Threading: take a database of known structures, take a training set of sequences s_i with known native structure, learn interaction parameters ε_{ij} by asking that

$$E(s_i, Nat(s_i)|\varepsilon_{ij}) << E(s_i, Wrong(s_i)|\varepsilon_{ij})$$

for all structures and sequences in the training set.



$$U = \sum_j \sum_{i<j} U_{SC_iSC_j} + w_{SC_p} \sum_j \sum_{i \neq j} U_{SC_iP_j} + w_{el} \sum_j \sum_{i<j-1} U_{P_iP_j}$$

$$+ w_{tor} \sum_i U_{tor}(\gamma_i) + w_{tord} \sum_i U_{tord}(\gamma_i, \gamma_{i+1})$$

$$+ w_b \sum_i U_b(\theta_i) + w_{rot} \sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i})$$

$$+ \sum_{m=2}^{N_{corr}} w_{corr}^{(m)} U_{corr}^{(m)} + w_{vib} \sum_i U_{vib}(d_i)$$

Fig. 1. UNRES model of the polypeptide chain. Filled circles represent p groups, and open circles represent the C^α atoms, which serve as geometric points. Ellipsoids represent side chains, with their centers of mass at the SCs. The p groups are located halfway between two consecutive C^α vectors or dCs. The SCs are located at the end of the C^α···SC vectors or the dXs. The variables to change the conformation of the polypeptide chain are the virtual-bond angles θ, the virtual-bond dihedral angles γ, and the angles α_{SC} and β_{SC}, which define the location of a side chain with respect to the backbone.

H.A. Scheraga et al, J. Comput. Chem. 18, 849 (1997)

Problems with this approach:

1) Potential of mean force: sampling native states is not equivalent to sampling all the configuration space!! There is no guarantee that the force field will lead the protein to the folded state.

2) Threading: dimensions of the training set, impossibility to fulfill all the inequalities
Results are generally not wonderful for proteins not in the training set.

However:
1) refinement can always be added
2) best performing ab-initio method in CASP resorts to realistic potential + inferred potentials (Baker's group)

Simple Models

Let's follow a completely different approach: let's try to find the minimal ingredients that produce a protein-like behavior.

- We will address questions like:

- what is the minimal number of species required?

- is protein-like behavior a property of a random sequence?

- How is protein-like behavior related to the structure of the energy landscape?

- Can we characterize the folding mechanisms and bottleneck?

Let's start by observing that
a hydrophobic homopolymer is not a protein!

A homopolymer has

- theta transition of II order from a self avoiding chain to a compact chain

- extensive number of compact states $M \propto (q/e)^N$: we cannot identify the native state of a protein with the compact phase of an homopolymer

- no secondary structure.

Homopolymer collapse can be a good model of just the first stages of folding (perhaps...)

Heteropolymers:

- can we say something in general?
- Is protein-like behavior a generic property?

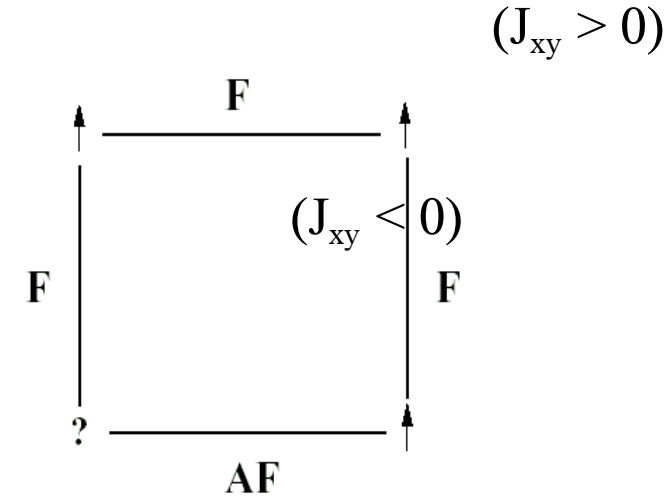
Digression: Complex (disordered) systems

- Example: spin glasses

- frustration: ground-state highly degenerate, many different structures

- properties depend on details of the hamiltonian.
  Some properties are self-averaging

- new techniques in statistical mechanics



$$\mathcal{H}_{\mathcal{J}} = -\sum_{<x,y>} J_{xy}\sigma_x\sigma_y - h\sum_x \sigma_x$$

Example of a complex, frustrated system: spin glass. In ?, both $\sigma = 1$ and $\sigma = -1$ are bad choices. Ground state is highly degenerate.

Random bond model on a cubic lattice

(Thirumalai,Klimov, cond-mat/0101048)

Consider a polymer of N residues, with energy:

$$E(\{r_i\}) = \sum_{i<j} \Delta(\mid r_i - r_j \mid -a) B_{ij}$$

where Δ() is 1 if two non-consecutive residues are next-neighbors on the lattice (of spacing "a"), and the interactions B_{ij} are random variables with distribution:

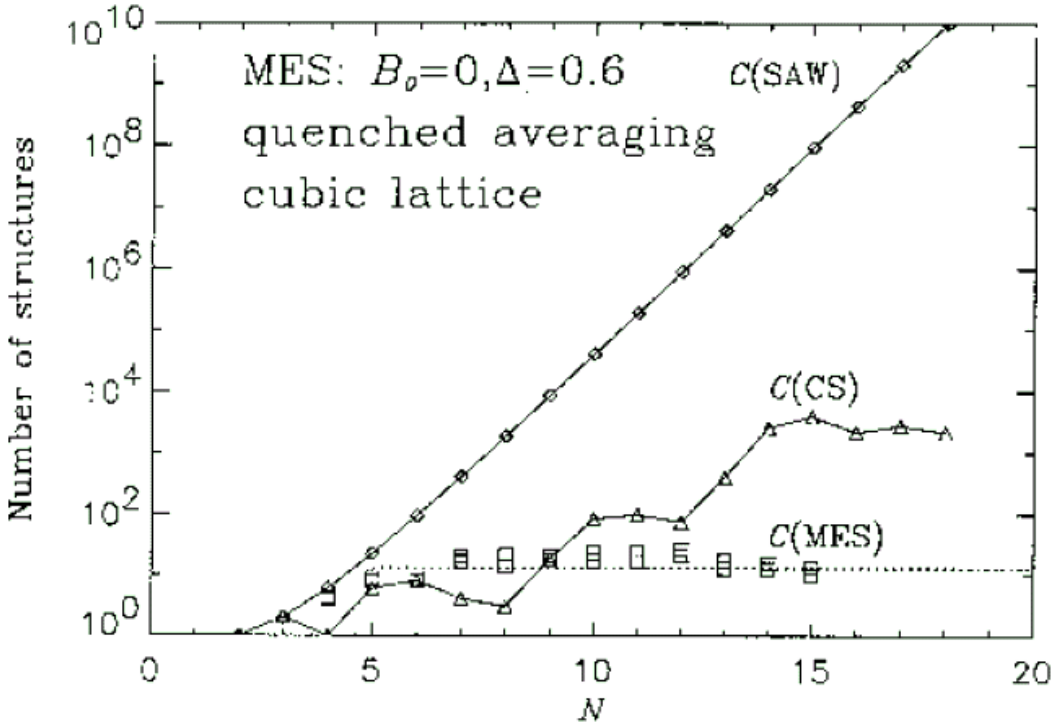
$$P(B_{ij}) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(B_{ij} - B_0)^2}{2\sigma^2}]$$

Implicit idea: identical pairs of residues (i,j), (i',j') at different positions in the chain will interact differently.

Exhaustively count SAW, Compact Structures (CS), Minimal Energy Compact Structures (MES) for N<=15 on a cubic lattice.

MES defined by E(MES)<Emin+ΔE, $C(\overline{\mathrm{MES}}) = exp[\overline{ln[c(\mathrm{MES})]}]$

B0=-0.1 (reproducing 55% hydrophobic, 45% hydrophilic as in proteins)
ΔE=0.6



The idea is that protein-like structures must be low lying ones, among all the possible structures that can be formed.
The use of a lattice model and of random bonds allows to infer "generic" properties of MES, independent of the realization of disorder.

Results:
- number of SAW and CS: exponential in N (as expected)
- number of MES: increases as log(N) (at most).

With minimal requirements of low energy and compactness, configuration space becomes sparse.

This suggests that many sequences will necessarily have the same native structure (but here we haven't sequences, so this is not a proof). More on this later on, with HP models.

Random Heteropolymers:
can we say something without lattice and exhaustive enumerations?

Let's see an analytic approach (Garel, Orland, Pitard, cond-mat/9706125):

- Very coarse model for the polymer

$$Z(\{v_{ij}\}) = \int \prod_i d\vec{r}_i \prod_i g(\vec{r}_i, \vec{r}_{i+1}) \ \exp(-\beta \mathcal{H}(\{v_{ij}\}))$$

- where $g(\vec{r}_i, \vec{r}_{i+1}) = \delta(|\vec{r}_i - \vec{r}_{i+1}| - a)$ or

- $$g(\vec{r}_i, \vec{r}_{i+1}) \to \exp(-\frac{d}{2a^2} \left(\frac{d\vec{r}(s)}{ds}\right)^2)$$

- Reduced Hamiltonian

$$\beta\mathcal{H}(\{v_{ij}\}) = \frac{1}{2}\sum_{i\neq j} v_{ij}(\vec{r}_i, \vec{r}_j) + \frac{1}{6}\sum_{i\neq j\neq k} w_0(\vec{r}_i, \vec{r}_j, \vec{r}_k) + \ldots$$

- The free energy is given by

$$F(\{v_{ij}\}) = -T\log Z(\{v_{ij}\})$$

- Assume 3-body term is not sequence dependent: excluded volume term

$$w_0(\vec{r}_i, \vec{r}_j, \vec{r}_k) = w\delta(\vec{r}_i - \vec{r}_j)\delta(\vec{r}_j - \vec{r}_k)$$

- Assume short-range sequence dependent two-body interaction:

$$v_{ij}(\vec{r}_i, \vec{r}_j) = v_{ij}\delta(\vec{r}_i - \vec{r}_j)$$

- To get an idea of the behavior of the system, assume quenched disorder:

$$F = \overline{F} = \int \prod P(\{v_{ij}\})F(\{v_{ij}\})d(\{v_{ij}\})$$

- "Average" heteropolymer

The random-bond heteropolymer

- It is defined by

$$v_{ij} = v_0 + w_{ij}$$

with w_{ij} random independent variables with distribution

$$h(w_{ij}) = \frac{1}{\sqrt{2\pi w^2}} \exp\left(-\frac{w_{ij}^2}{2w^2}\right)$$

- Average over Replicas

$$\overline{Z^n} = \int \prod_a \mathcal{D}\vec{r}_i^a \prod_{i,a} g(\vec{r}_i^a, \vec{r}_{i+1}^a) \exp\left(-\tilde{v}_0 \sum_{i<j}\sum_a \delta(\vec{r}_i^a - \vec{r}_j^a) + \frac{\beta^2 w^2}{2} \sum_{a\neq b}\sum_{i<j} \delta(\vec{r}_i^a - \vec{r}_j^a)\delta(\vec{r}_i^b - \vec{r}_j^b)\right)$$

$$\times \exp\left(-\frac{w_0}{6} \sum_{i\neq j\neq k}\sum_a \delta(\vec{r}_i^a - \vec{r}_j^a)\, \delta(\vec{r}_j^a - \vec{r}_k^a)\right)$$

with $\tilde{v}_0 = v_0 - \beta^2 \frac{w^2}{2}$.

- Introduce density order parameters and overlap order parameters (and conjugate)

$$\overline{Z^n} = \int \mathcal{D}q_{ab}(\vec{r},\vec{r}')\mathcal{D}\hat{q}_{ab}(\vec{r},\vec{r}')\mathcal{D}\rho_a(\vec{r})\mathcal{D}\phi_a(\vec{r}) \exp\left(G(q_{ab},\hat{q}_{ab},\rho_a,\phi_a) + \log\zeta(\hat{q}_{ab},\phi_a)\right)$$

$$\rho_\alpha(\vec{r}) = \sum_{i=1}^N \delta(\vec{r}_i^{(\alpha)} - \vec{r})$$

$$q_{\alpha,\beta}(\vec{r},\vec{r}') = \sum_{i=1}^N \delta(\vec{r}_i^{(\alpha)} - \vec{r})\delta(\vec{r}_i^{(\beta)} - \vec{r}')$$

$$G(q_{ab}, \hat{q}_{ab}, \rho_a, \phi_a) = \int d^d r \sum_a \left(i\rho_a(\vec{r})\phi_a(\vec{r}) - (\tilde{v}_0)\ \frac{\rho_a^2(\vec{r})}{2} - \frac{w_0}{6}\rho_a^3(\vec{r}) \right)$$
$$+ \int d^d r \int d^d r' \sum_{a<b} \left(i q_{ab}(\vec{r}, \vec{r}')\hat{q}_{ab}(\vec{r}, \vec{r}') + \frac{\beta^2 w^2}{2}\ q_{ab}^2(\vec{r}, \vec{r}') \right)$$

$$\zeta(\hat{q}_{ab}, \phi_a) = \int \mathcal{D}\vec{r}_a(s) \exp\left(-\frac{d}{2a^2} \int_0^N ds\ \dot{\vec{r}}_a^{\,2} \right)$$
$$\times\ \exp\left(-i \int_0^N ds\ \sum_a \phi_a(\vec{r}_a(s)) - i \int_0^N ds\ \sum_{a<b} \hat{q}_{ab}(\vec{r}_a(s), \vec{r}_b(s)) \right)$$

- at high temperature, $\rho_\alpha = 0$ and $q_{\alpha,\beta} = 0$
  It is the Denatured phase (SAW)

- at lower temperature, T_θ , collapsed phase
  with $\rho_\alpha \neq 0$ and $q_{\alpha,\beta} = 0$ Molten Globule

- at lower temperature, T_F , freezing
  transition $\rho_\alpha \neq 0$ and $q_{\alpha,\beta} \neq 0$ Native?

Frozen state ≠ native.
What we get is that in the frozen phase we have that the number of possible configurations is not exponential any more.

But there is not unique native state, and low lying states can act as traps for the kinetics.

Let's see this with the simplest frustrated model: it can be proved indeed that the Random Bond Heteropolymer Model, when considering just compact conformations, is equivalent to the Random Energy Model, introduced for spin glasses, where chain connectivity is neglected.

Random Energy Model: application to proteins

(D.J.Wales, ”Energy Landscapes”, Cambridge Univ. Press,2003, 9.4; Bryngelson e al., Proteins: Struct.,Funct and Gen.,21,167,1995)

Let's take a biased approach: suppose that we have a “protein”, that we characterize as having a special, “native” state.

Let's introduce an order parameter: a measure “n” of the structural similarity to the native state, normalized between 0 and 1.

Let's suppose that for any value of ”n”, the energy of a state with that “n” is a gaussian random variable:

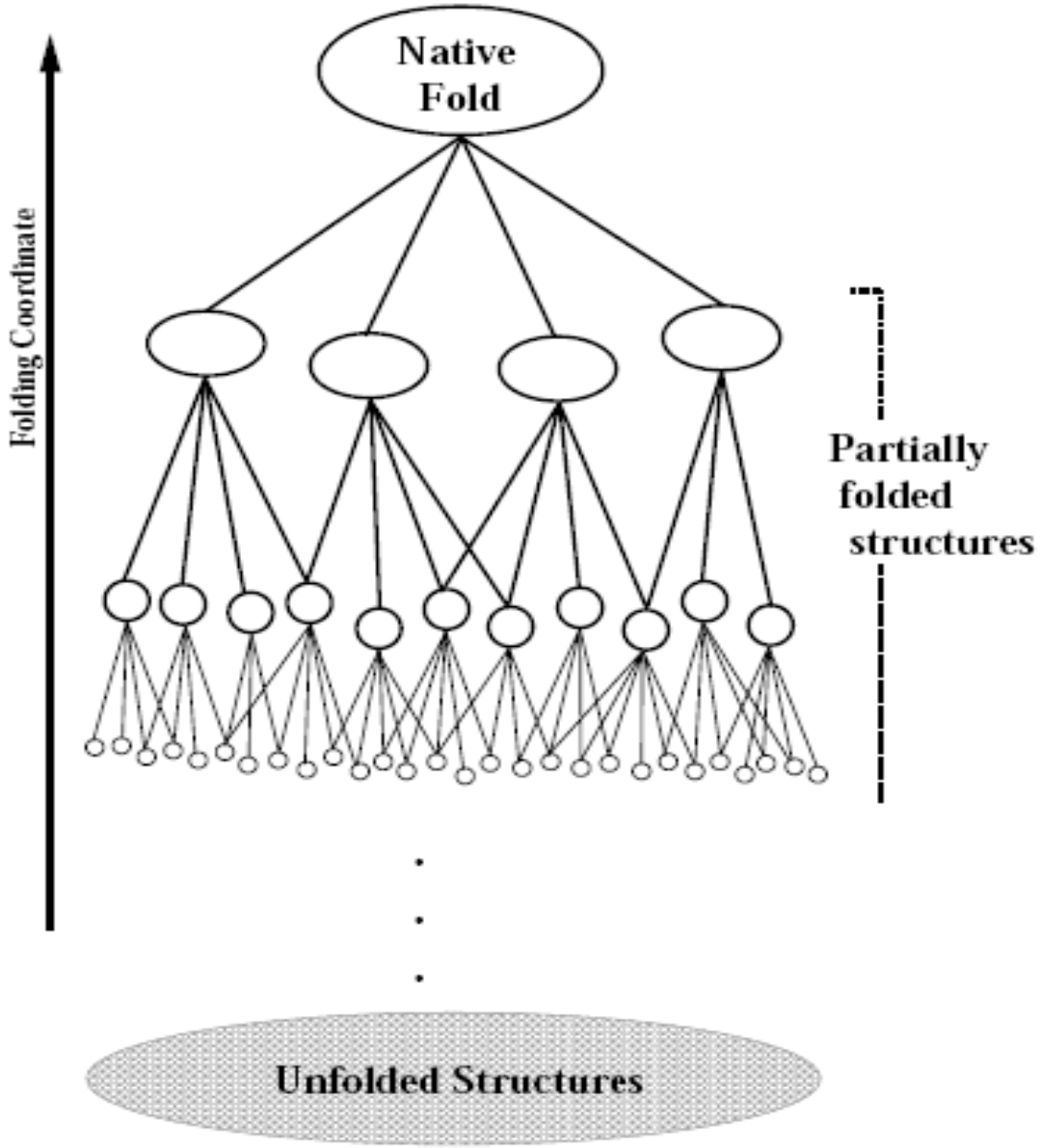
$$P(E,n)=\frac{1}{\sqrt{2\pi\Delta E^2(n)}}\exp\left(\frac{-[E-\bar{E}(n)]^2}{2\Delta E^2(n)}\right)$$

so that the number of states at a given “n” will be given by

$$\Omega(E,n)=\Omega(n)P(E,n)$$

(possible expression: $\Omega(n)\propto\gamma^{N(1-n)}$)

number of configurations at a certain “n” strongly dependent on “n”:



Let's put in the Boltzmann factor:

$$p(E,n) = \frac{1}{Z}\Omega(E,n)e^{-E/k_BT}$$

The most probable structure at fixed similarity “n” and at temperature T will be the maximum of p(E,n):

$$E_{m.p.}(n) = \bar{E}(n) - \frac{\Delta E(n)^2}{k_BT} \ ,$$

Assuming negligible fluctuations, the corresponding number of states and entropy will be:

$$\Omega(E_{m.p.}(n),n) = \exp\left[\frac{S_0(n)}{k_B} - \frac{\Delta E(n)^2}{2(k_BT)^2}\right]$$

$$S(E_{m.p.}(n),n) = k_B \log(\Omega(E_{m.p.}(n),n)).$$

whence one can estimate the free-energy as

$$\begin{aligned} F(n) &= E_{m.p.}(n) - TS(E_{m.p.}(n), n) \\ &= \overline{E}(n) - \frac{\Delta E(n)^2}{2k_B T} - TS_0(n) \ . \end{aligned}$$

This can have different forms (we will discuss them in the following): downhill, two-state,....

Important remark:

Notice that the expression of $S(E_{m.p.}(n), n)$ vanishes at $T=T_g(n)$:

$$T_g(n) = \left(\frac{\Delta E(n)^2}{2k_B S_0(n)} \right)^{1/2}$$ glass transition temperature

where the system "runs out" of entropy, and can live in one of few states, that depends on the realization of disorder.

What kind of kinetics emerges on such a landscape?

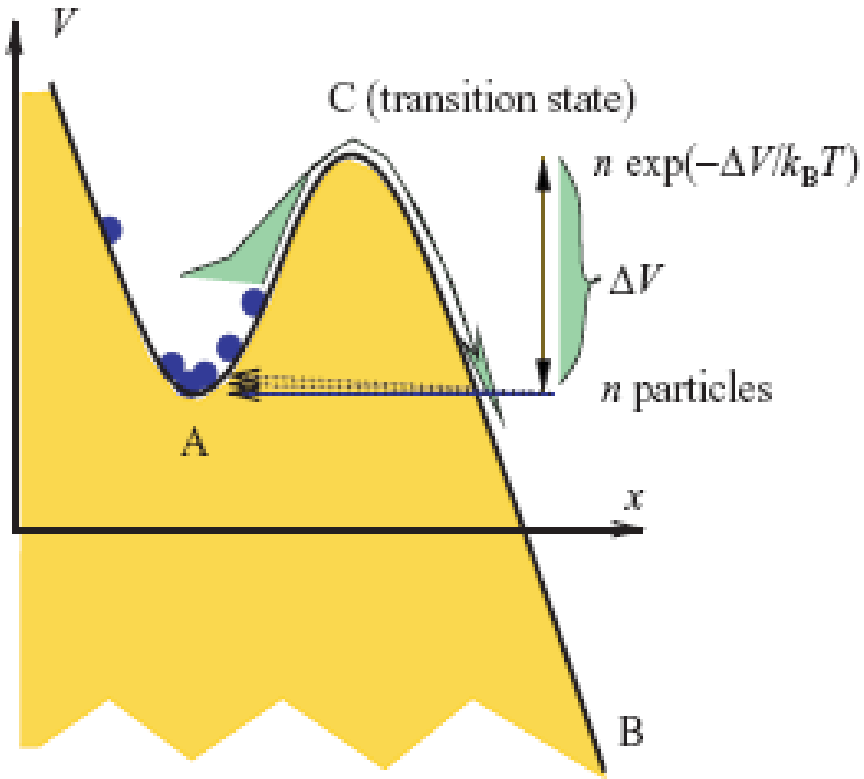
Digression: Kramer's theory (overdamped)

Given a one-dimensional system, with two states separated by a barrier, and assuming stationary conditions, with negligible current, the rate for a particle in A to cross the barrier is

$$r=\frac{\omega_A \omega_C \tau}{2\pi}\exp\left(-\frac{(V_C-V_A)}{k_B T}\right)$$

where ω_A , ω_C represent the widths of the well A and the transition state.

$\tau=mD/(k_B T)$, with D the diffusion coefficient along the reaction coordinate



Is this discussion still valid for REM?

Kramer's formula is derived under the assumptions:

1) we are dealing with just one path: well defined reaction coordinate
2) no configurational entropy in the various states
3) the pre-exponential coefficient accounts for the shape of the potential and for the diffusion coefficient, which is homogeneous along the reaction coordinate. Solvent is just a source of friction

What we have here in our REM model:

1) in principle, many pathways between states at different “n”
2) entropy
3) diffusion is related to the escape time from the states at given “n”: we will see that this is dependent on the coordinate “n”

REM kinetics:

the expression of $E_{m.p.}$ tells us that the most likely conformation at T does not coincide with the peak of the distribution.

$$E_{m.p.}(T,n)=\bar{E}(n)-\frac{\Delta E^2(n)}{k_B T}$$

That is, there is a gap between the most likely and the majority of its neighbors.

This implies that, there will be a typical "microscopical" escape time with a super-Arrhenius behavior (Ferry law)

$$\bar{t}(n) = t_0 e^{(\Delta E(n)/k_B T)^2}$$

NB: this holds true out of the frozen phase, where we can talk of "typical" behavior! Indeed, this assumes that there are many "channels" out of the most likely state. But this is not always true!!

In the frozen phase:

When $T<=T_g(n)$, with

$$T_g(n) = \left(\frac{\Delta E(n)^2}{2k_B S_0(n)} \right)^{1/2}$$

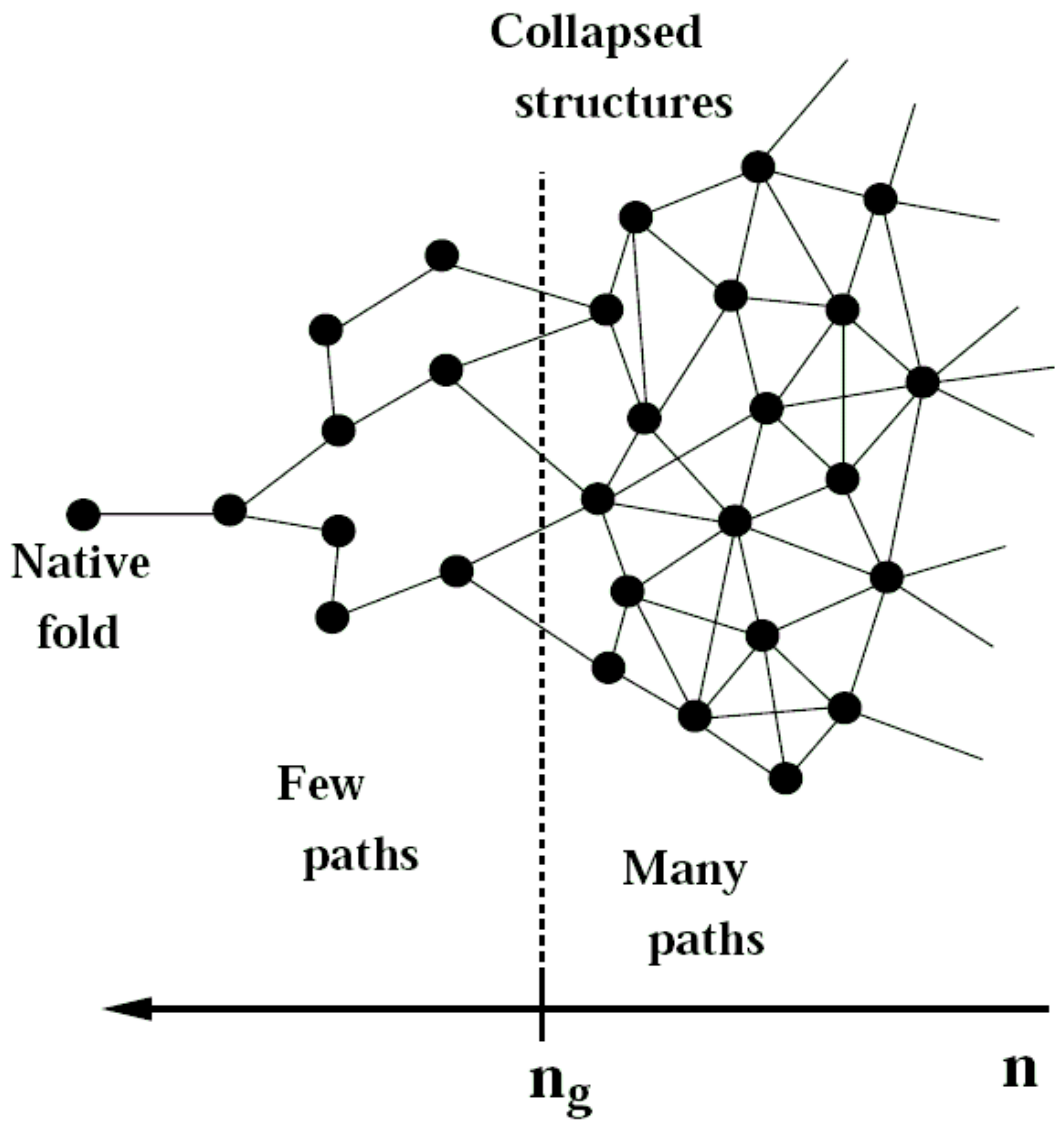
the escape time will need a random search of the one, dominant channel, in a "Levinthal" way, which will take no shorter than

$$\bar{t} = t_0 e^{S_0(n)/k_B}$$

Notice that the freezing can also be reached at fixed temperature, upon increasing n:

(at fixed n, $T<T_g$) or (at fixed T, $n>n_g$) --> frozen phase.

There is a sudden fall in the number of paths if freezing takes place...

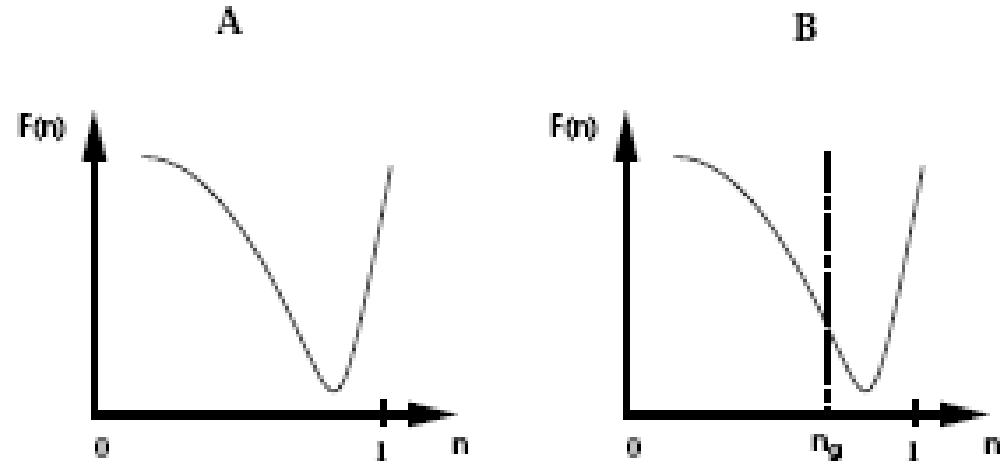


Kinetics scenarios:

0) Downhill, with or without freezing
Folding time= maximum of

$$\tau = \bar{t}(n^{\ddagger}_{kin}).$$

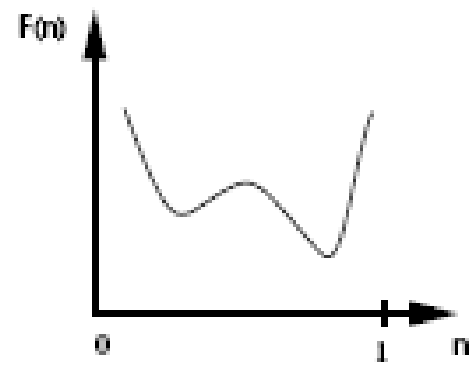
Type 0



1) two-state, no freezing
2) two-state, late or early freezing.
Folding time= maximum of

$$\tau = \bar{t}(n^{\ddagger}_{kin})e^{F^{\ddagger}_{kin}/k_BT}$$

Type I

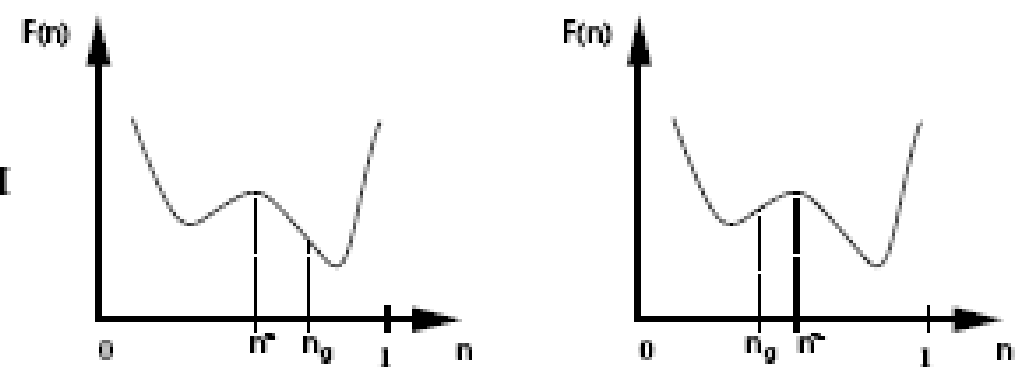


$n^{\ddagger}_{kin}$ =the value minimizing τ above

$$F^{\ddagger}_{kin} = F(n^{\ddagger}_{kin}) - F(n_{uf})$$

$$\log\left(\frac{\tau}{t_0}\right) = \frac{F^{\ddagger}_{kin}}{k_BT} + \frac{\Delta E(n^{\ddagger}_{kin})^2}{(k_BT)^2}$$

Type II



Again, folding time is not Arrhenius-like (exp(barrier/T))

Important remarks:

- Type I scenario is most likely to be the relevant one for protein folding.

- Type 0 scenario could appear from a type I at low or high temperatures. Recent (but controversial) claims of experimental observation of a protein with a type 0 (downhill) scenario at all temperature

- In every case, scenarios with a freezing transition, especially an early one, are unlikely for proteins, because they wouldn't be robust against mutations: in the frozen region, minimal energy configurations and folding paths depend on the realization of disorder, i.e. on sequence, and a mutation could change completely the overall pathway.

Experimentally, some similarities and some differences:
ϕ-value analysis says that most mutations don't change the path, but some probably do. Probably folding paths are few and quite degenerate on sequences. (More on this ideas later on, with microscopic characterization and Go-models.)

Can we trust these results? let's revise the hypotheses:

1) We are working with a (modified version of a) REM which has been proved to be equivalent to the compact phase of a heteropolymer model.
So when we talk of type 0,I scenarios we are referring just to compact states, without any control on the geometry of the collapsing chain, and on the coupling between the latter and the similarity to the native state “n”

2) existence of a reaction coordinate “n”: folding proceeds along the gradient of F(n).
Very strong assumption: we are projecting the dynamics on just one coordinate, “n”, and are acting as if motion along “n” was slow, with all other degrees of freedom at equilibrium at given “n”.
But actually we don't know if this is true!

3) at any “n”, gaussian distribution of the energies
Probably reasonable at low n, could break down at high n (close to native).
Moreover, in calculating F just using $E_{m.p.}$ we ignored fluctuations, implicitly assuming that there is a huge number of states at any n (no freezing).

4) when estimating barriers, we assumed that the close-neighbors of the most likely configuration at a given “n” are still configurations of the same “n”.
This is related to 2) and presents the same problems.

In every case, REM predicts relaxation times costrained by the escape from the huge number of minima of a landscape, characterized by typical barriers $\Delta E^2(n)/(k_B T)$.

So, according to REM:
a generic polypeptide is expected to have a complex (glassy-like) behavior: frustration, rugged energy landscape, glass transition, slow dynamics,...

This would be in general incompatible with function, and hence, life. Evolution has optimized sequences/structures for producing quick and stable folders.

The landscape must be “tuned”, in the sense that the roughness must be small, and an overall driving force towards native state must be present... Can we make this statement quantitative?

Let's consider a two-state system, with minima at n_U, n_N :
assuming $F(n_N) \approx \bar{E}(n_N)$, we can define the mid-folding temperature T_f (when $F(n_N)=F(n_U)$) and find:

$$\frac{T_f}{T_g(n_U)} = \frac{1}{y}\left(1+\sqrt{(1-y^2)}\right)$$

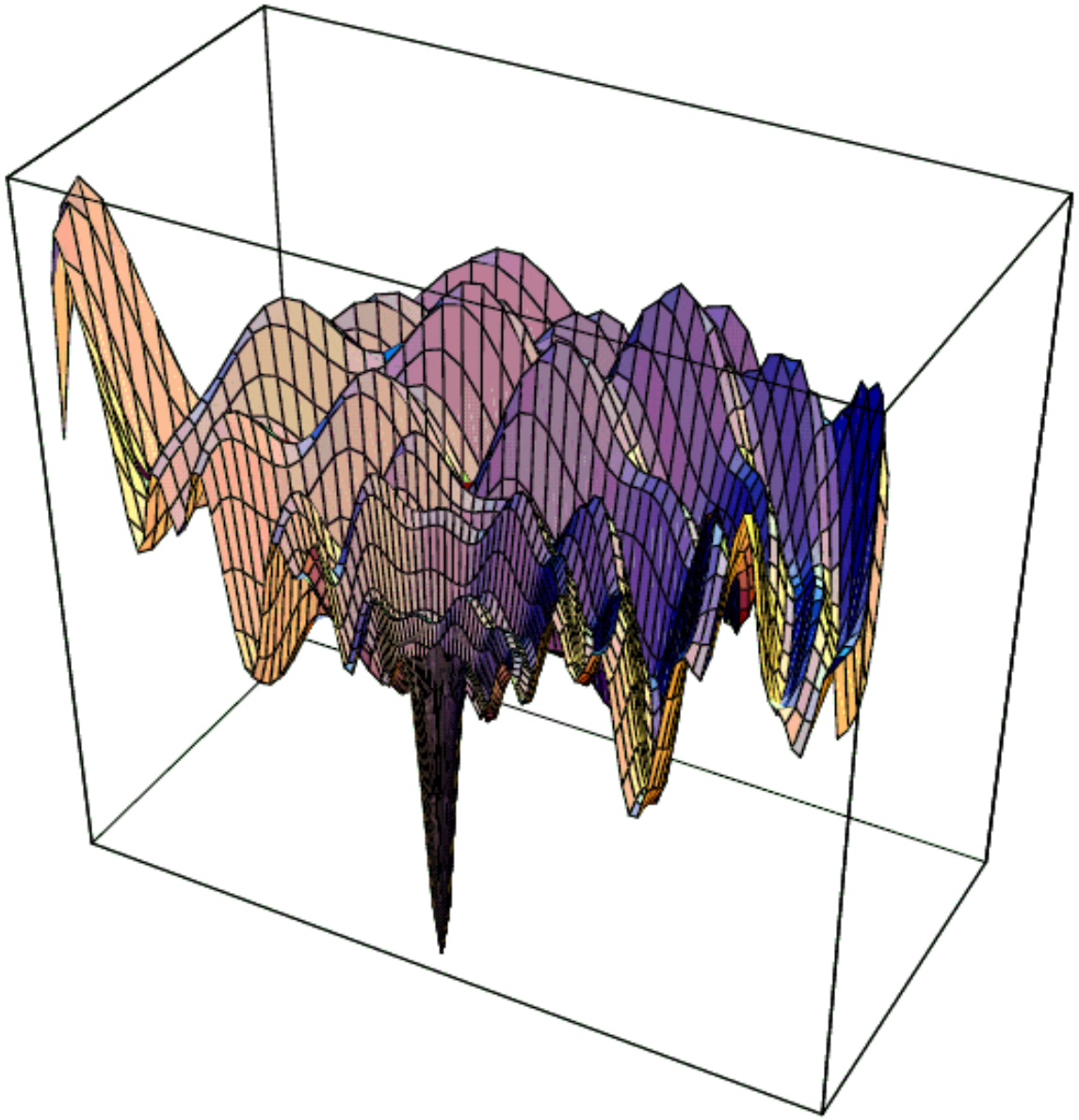
with

$$y = \sqrt{\frac{2S_0(n_U)}{k_B}}\left(\frac{\Delta E(n_U)}{\bar{E}(n_U)-\bar{E}(n_N)}\right)$$

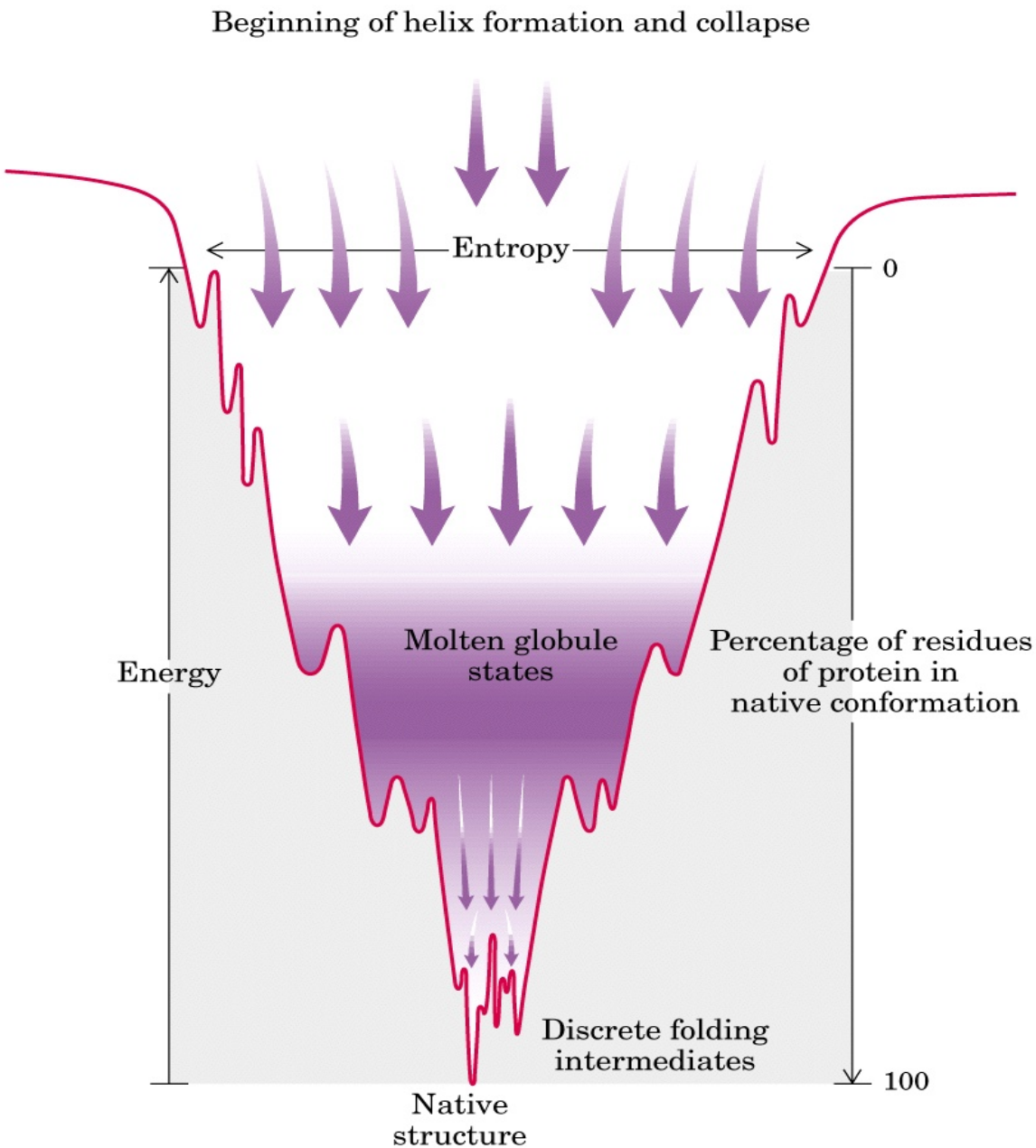
Fast folding has been associated to maximizing T_f/T_g, which corresponds to minimizing ratio of the "roughness" ΔE over the stability ($\bar{E}(n_U)$-$\bar{E}(n_N)$).
Fast folding implies a reasonably smooth landscape, where barriers are small compared to the stability gap!

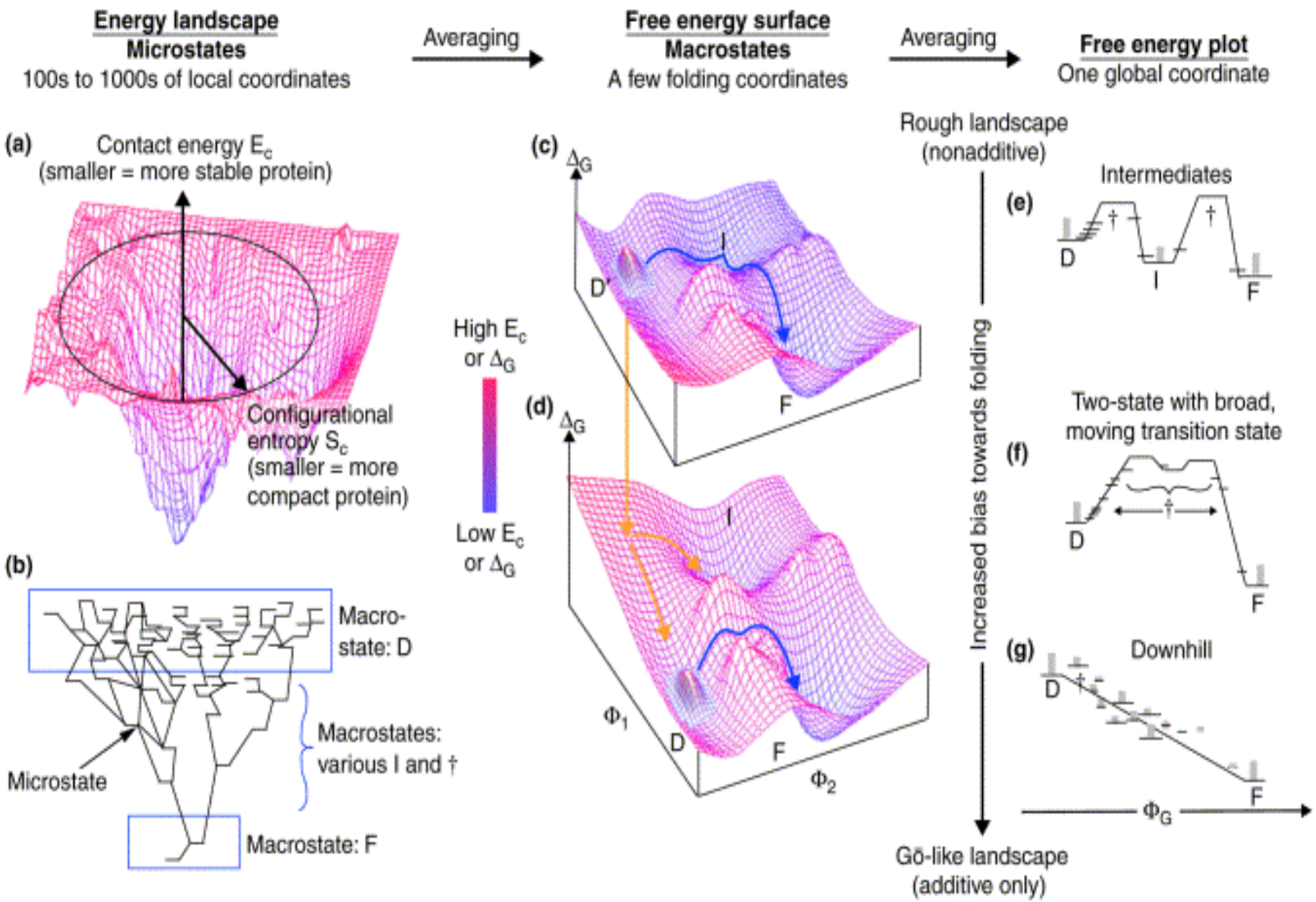
Exercise: following the lines of the above discussion of REM, prove the above result.

...A moderately rough landscape...



A good protein must present a reasonably smooth “folding funnel”!





Current Opinion in Structural Biology

Current understanding (from experiments and theory):

Hydrophobicity main driving "force" to collapse and important for stability, but correct folding is the product of many interactions
Stability is a result of huge cancellations of terms

A principle of minimal frustration is at work:

- native state= free energy minimum.

- no structurally different competing states at similar energy

- small energy barriers compared to driving force
  (little frustration: small ΔE/(stability gap))

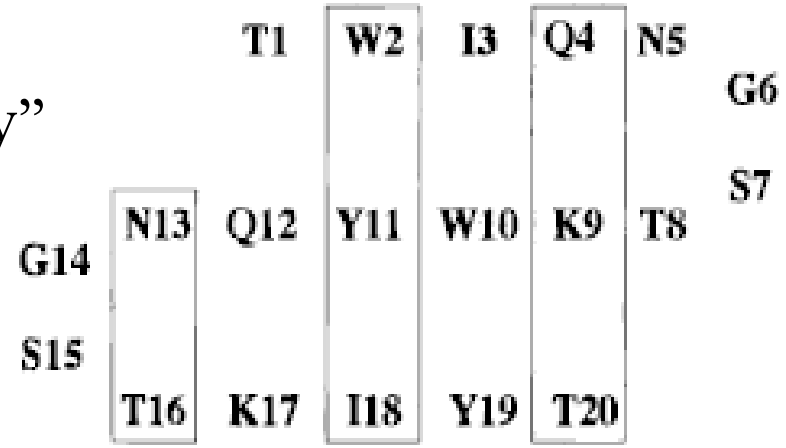
- huge funnel reduces the search in configurational space
  (NS must be reached by "all" initial conditions)

- multiple pathways, transition state ensemble, entropic barriers

Let's have a “microscopic look” into the configuration space...
(Rao and Caflish, JMB 2004)

- characterize the “minima” and “connectivity” of small designed peptide beta3s.

- nodes=“minima” and links=”paths” are identified dynamically:



CHARMM-PARAM19 (all heavy atoms, H bound to N or O).
Implicit water. 72 foldings + 73 unfoldings; average folding time = 83 ns.
$5x10^5$ snapshots (every 20 ps). T=330K

“Conformation”= a string of sec structure: e.g -EEEESSEEEEEES SEEEE-
Weight “w”= the frequency of a conformation
Found 1287 “nodes” with significant weight (w>20).

“link” between nodes that are visited within 20 ps or that are separated by conformations with little weight.

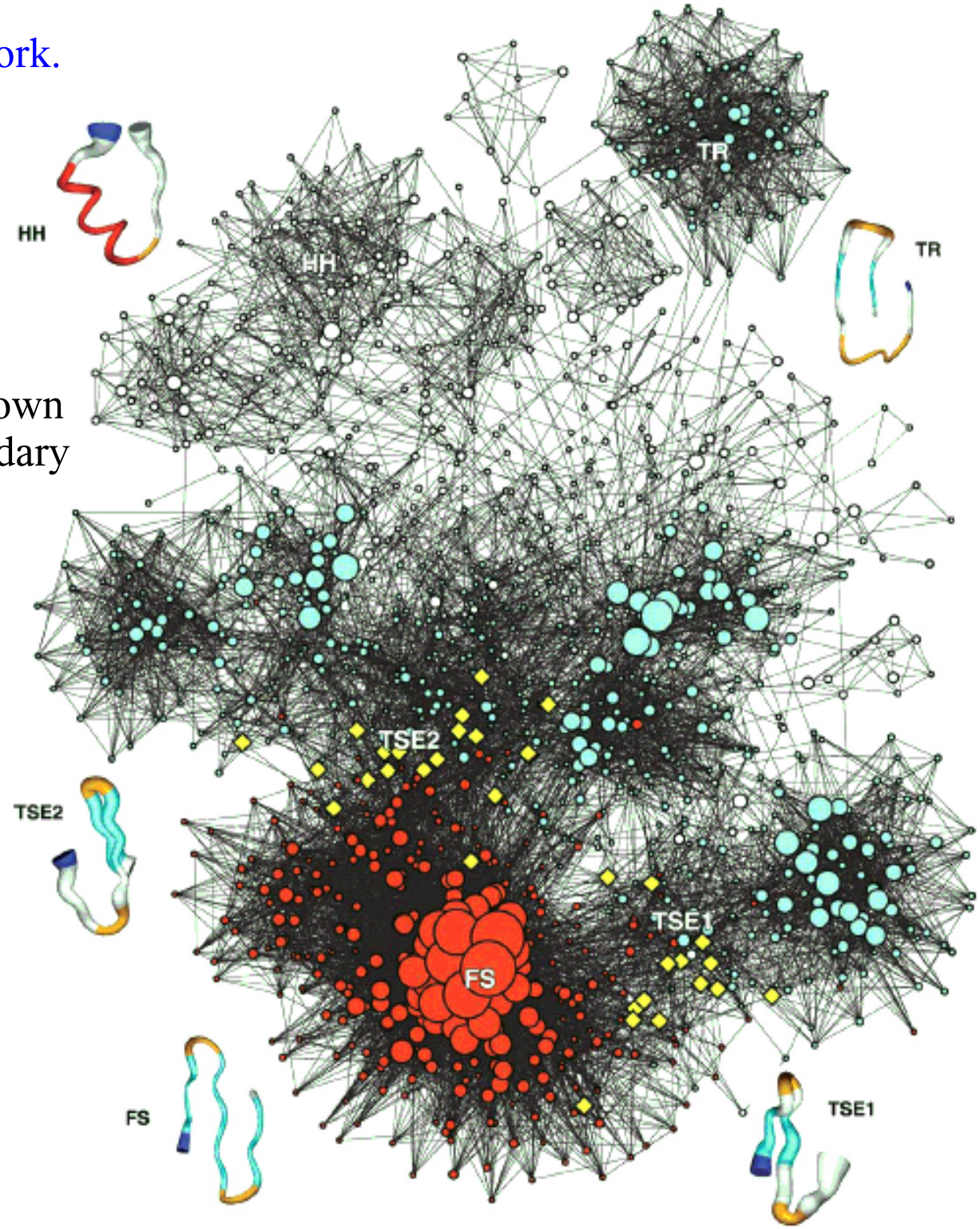
The beta3s conformation space network.

Node size = statistical weight w.
Node color= average neighbor connectivity k_{nn}: White $k_{nn}<30$; cyan: $30<k_{nn}<70$; red: $k_{nn}>70$.

Representative conformations are shown by a pipe colored according to secondary structure: white=coil; red=α-helix; orange=bend; cyan=strand; N terminus is blue.

The variable radius of the pipe reflects structural variability.

The yellow diamonds are folding TS conformations (TSE1, TSE2) characterized by a connectivity/weight ratio $k/2w>0.3$, a clustering coefficient $C<0.3$, and $60<knn<80$.



Remarks:
Difficult to find a good projection along a reaction coordinate.
Fraction of native contacts fails! Even the ensemble with half of the native contacts is heterogeneous and hard to classify.

Can we identify minima and transition states from a network analysis?

Tools:

k_{nn} = average neighbor connectivity, i.e. the average number of links of the neighbors of a given node.

C= average clustering coefficient: probability that any two neighbors of a node are connected.

The TS conformations are saddle points, i.e. local maxima with respect to the reaction coordinate for folding and local minima with respect to all other coordinates. For this reason:
TS identified as the nodes with a high connectivity/weight ratio $ki/2w>0.3$ and low clustering C.

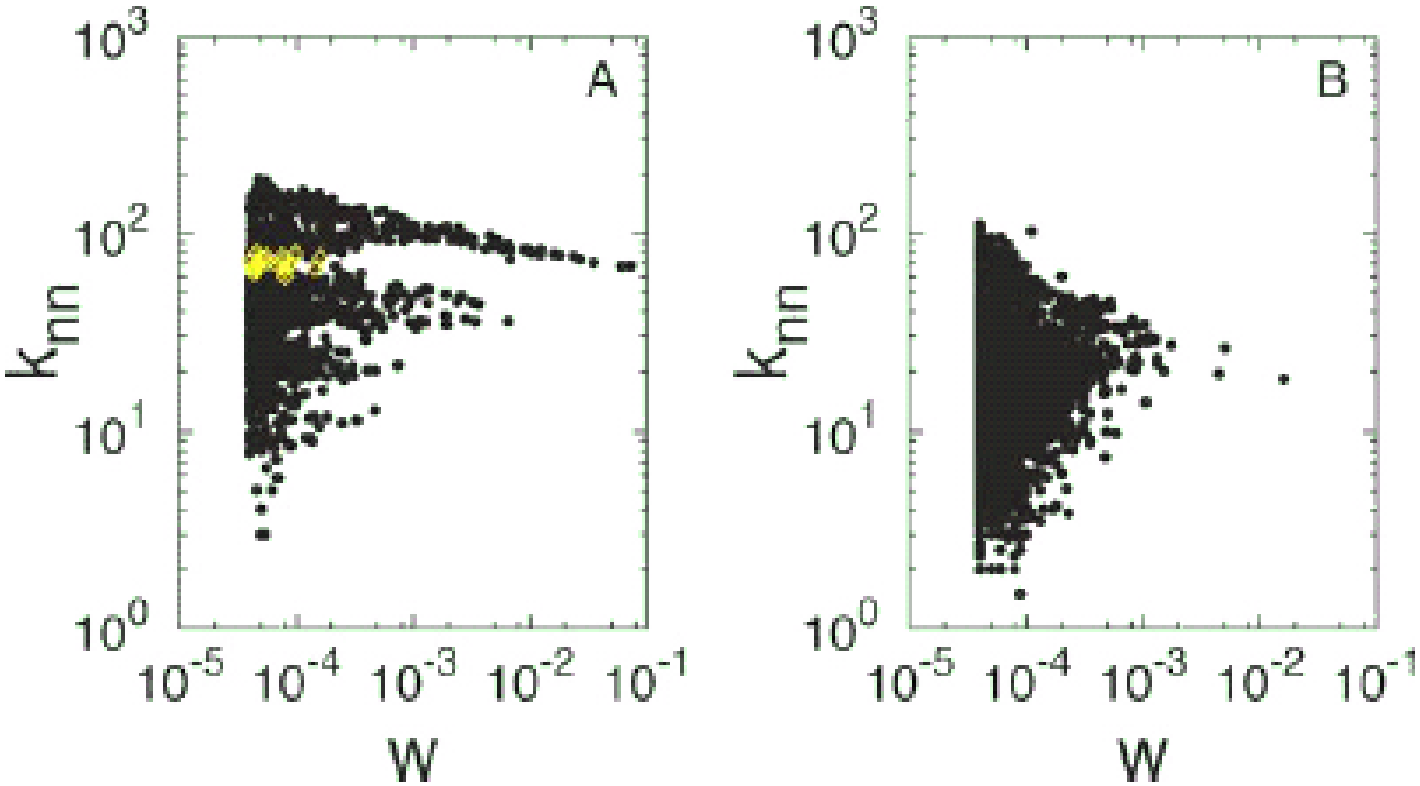
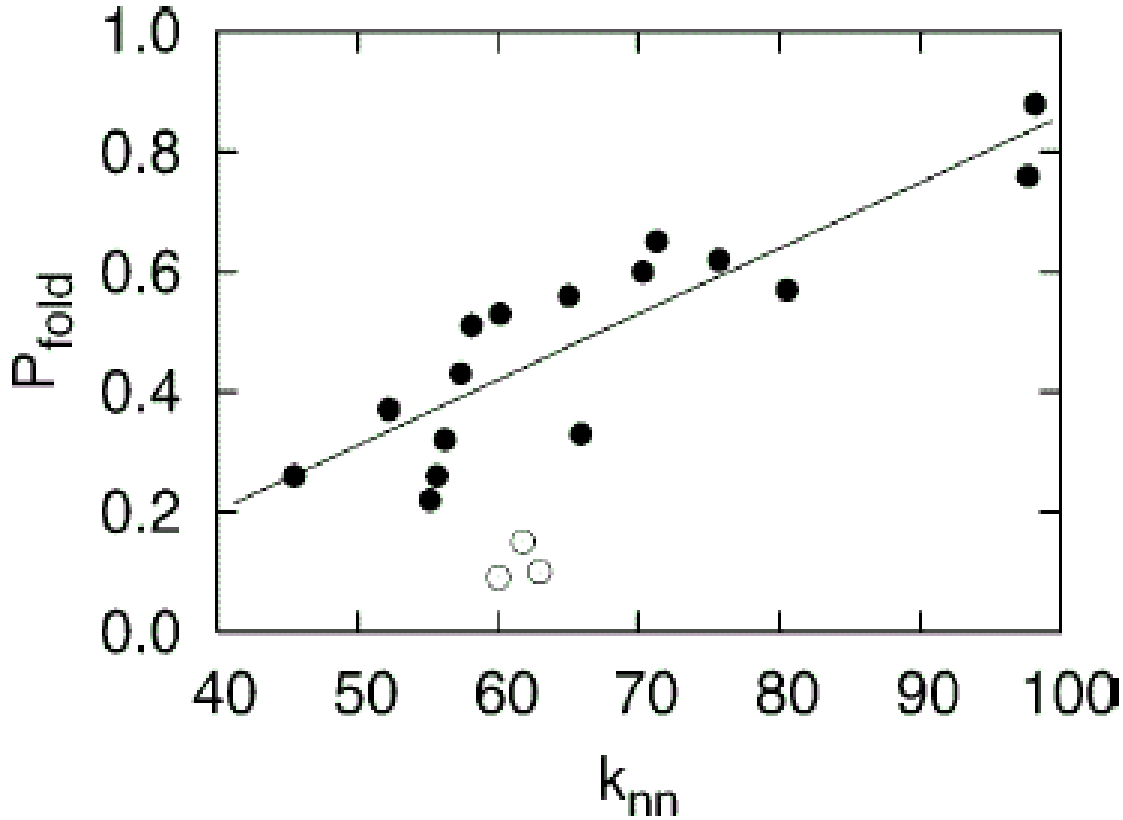


Figure 3. Average neighbor connectivity k_{nn} plotted as a function of the statistical weight for the 1287 nodes of beta3s (A) and for the 2658 nodes of the random heteropolymer (B). k_{nn} of node i is the average number of links of the neighbors of node i. The yellow diamonds are folding TS conformations (see also Figure 1 and the text) characterized by a connectivity/weight ratio $k/2\bar{w} > 0.3$, a clustering coefficient $C < 0.3$, and $60 < k_{nn} < 80$.



Correlation between P_{fold} and average neighbor connectivity k_{nn}. For nodes with high connectivity/weight ratio and low clustering coefficient, correlation = 0.89

Open circles: three nodes used as a negative control (low connectivity/weight ratio and/or high clustering coefficient but similar fraction of native contacts)

Observations:

- Conformation space network is scale-free (power-law on the degree distribution: $P(k) \sim k^{-\gamma}$). Also for the random heteropolymer.

- The most connected nodes are also low-lying minima on the free-energy landscape, yielding "basins" in the landscape.
The native basin of beta3s shows a hierarchical organization of conformations, not observed for the random heteropolymer.
Very heterogenous denatured state.

- C(beta3s)=0.49; C(random heteropolymer) = 0.28.
One order of magnitude larger than random networks with the same amount of nodes and links.(...polymeric nature?).

- Free energy minima and their connectivity emerge from the network analysis, without arbitrary reaction coordinates.
TS conformations: nodes with high k_{nn}/w ratio, low C. For TS conformations k_{nn} correlates with P_{fold}, the probability of folding.

A better (but even more painful) microscopic description:
The folding funnel in terms of disconnectivity graphs

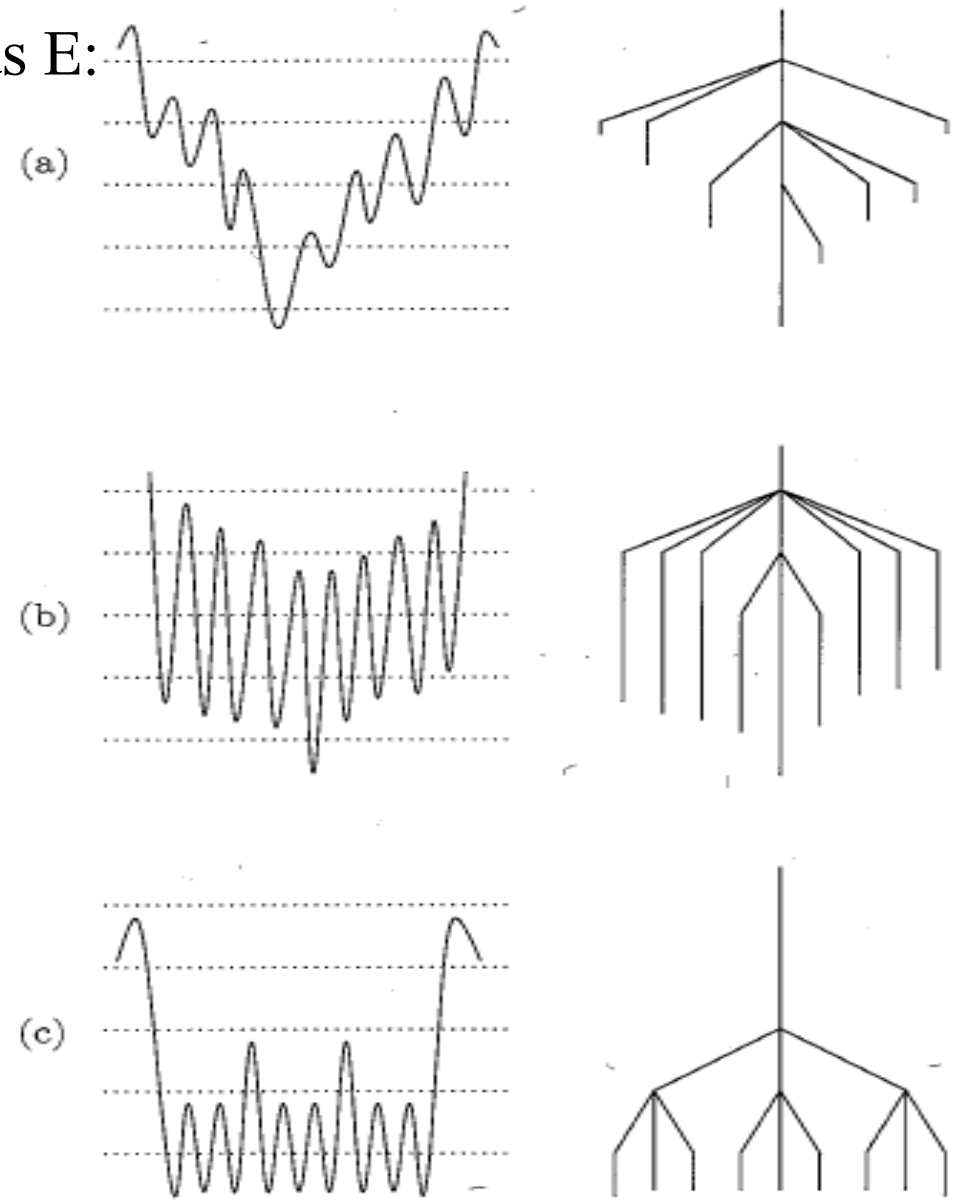
Cutoff at increasing energy thresholds E: minima becomes connected if the lowest barrier between them doesn't exceed E.

a) funneled landscape

b) still funneled, but high barriers

c) rough landscape: low interconversion from distant minima

Remarks: clear understanding of the "overall picture", but how can we find all the stationary states?



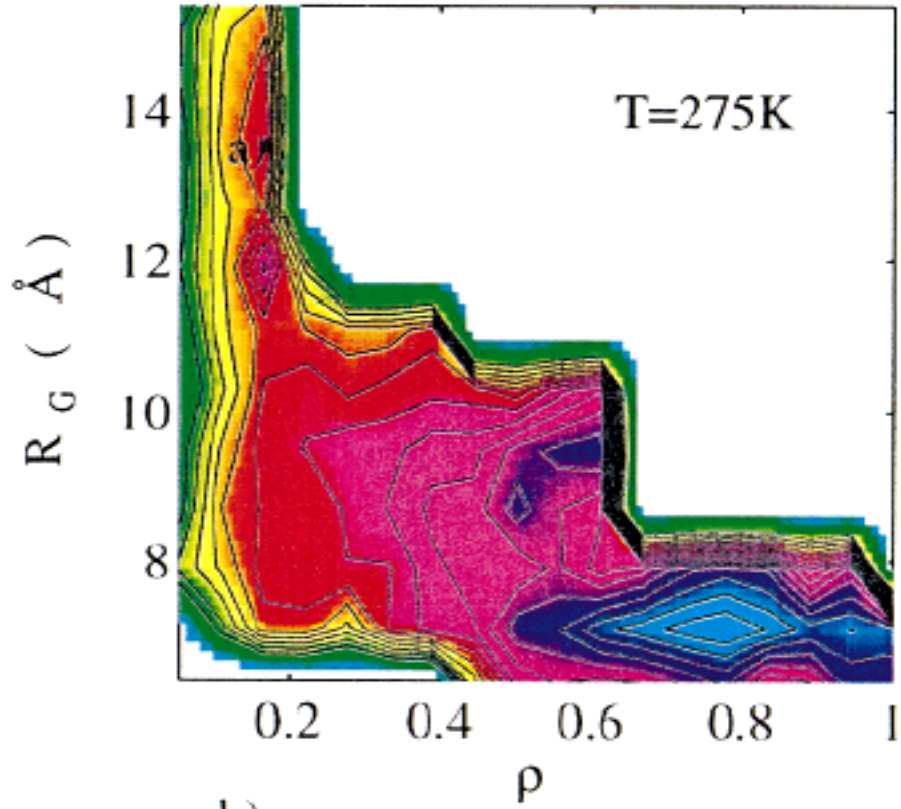
The detailed microscopic exploration of all the stationary points in the landscape is sound and interesting, but it is extremely difficult to perform.

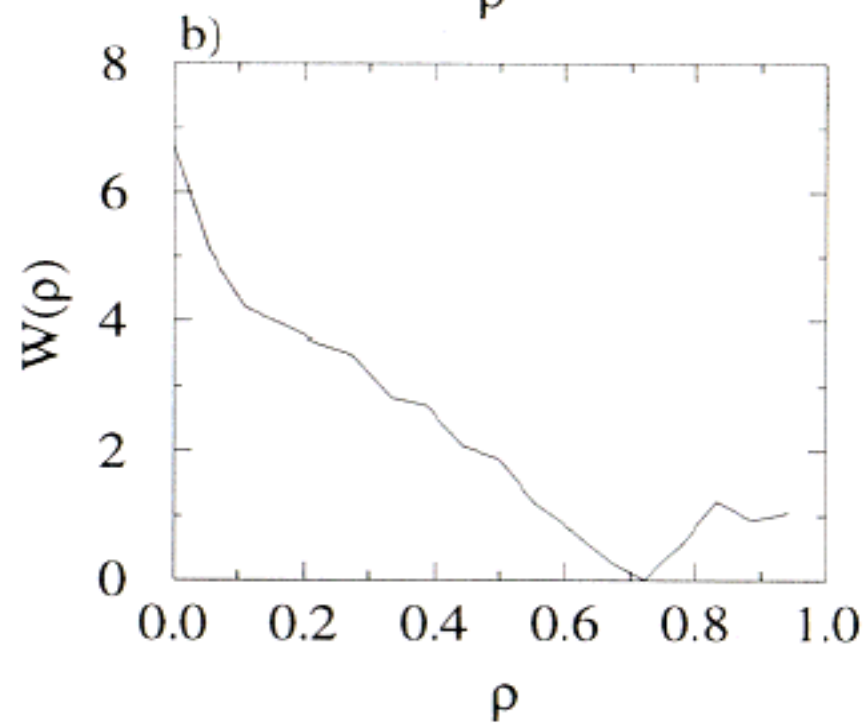
Indeed, resorting to projections on a few coordinates is often the best way to understand something, even in all-atoms simulations.

Remember though that the conclusions are as reliable as the choice of the reaction coordinate is...

...Indeed, not always the number of native contacts is a good coordinate...

Protein betanova, 20 residues.
CHARMM, explicit water





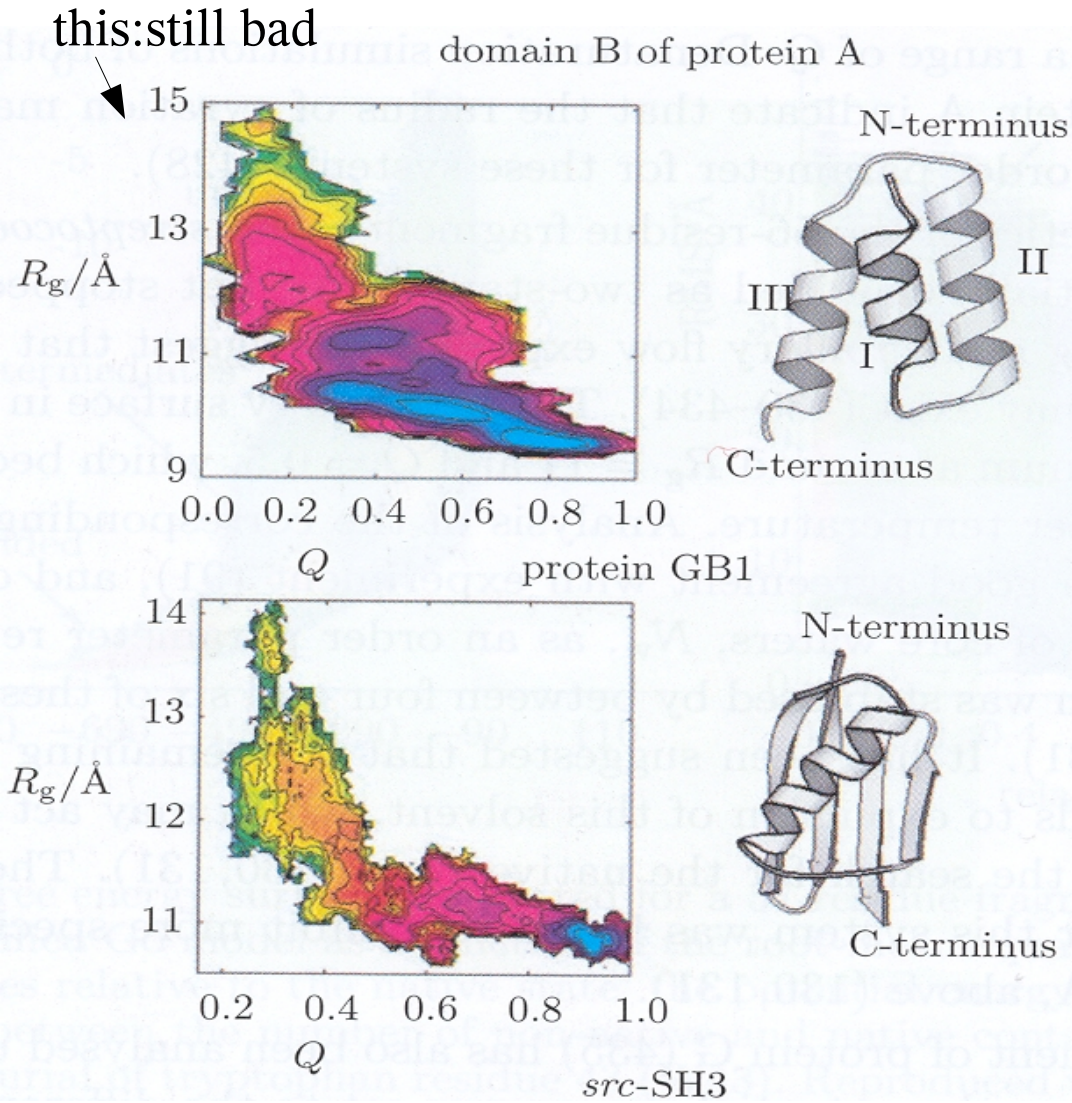
(a) Potential of mean force as a function of radius of gyration, Rg, and fraction of native contacts, ρ. The contours are drawn every 0.5 kcal/mol.
(b) Potential of mean force as a function of fraction of native contacts, ρ.

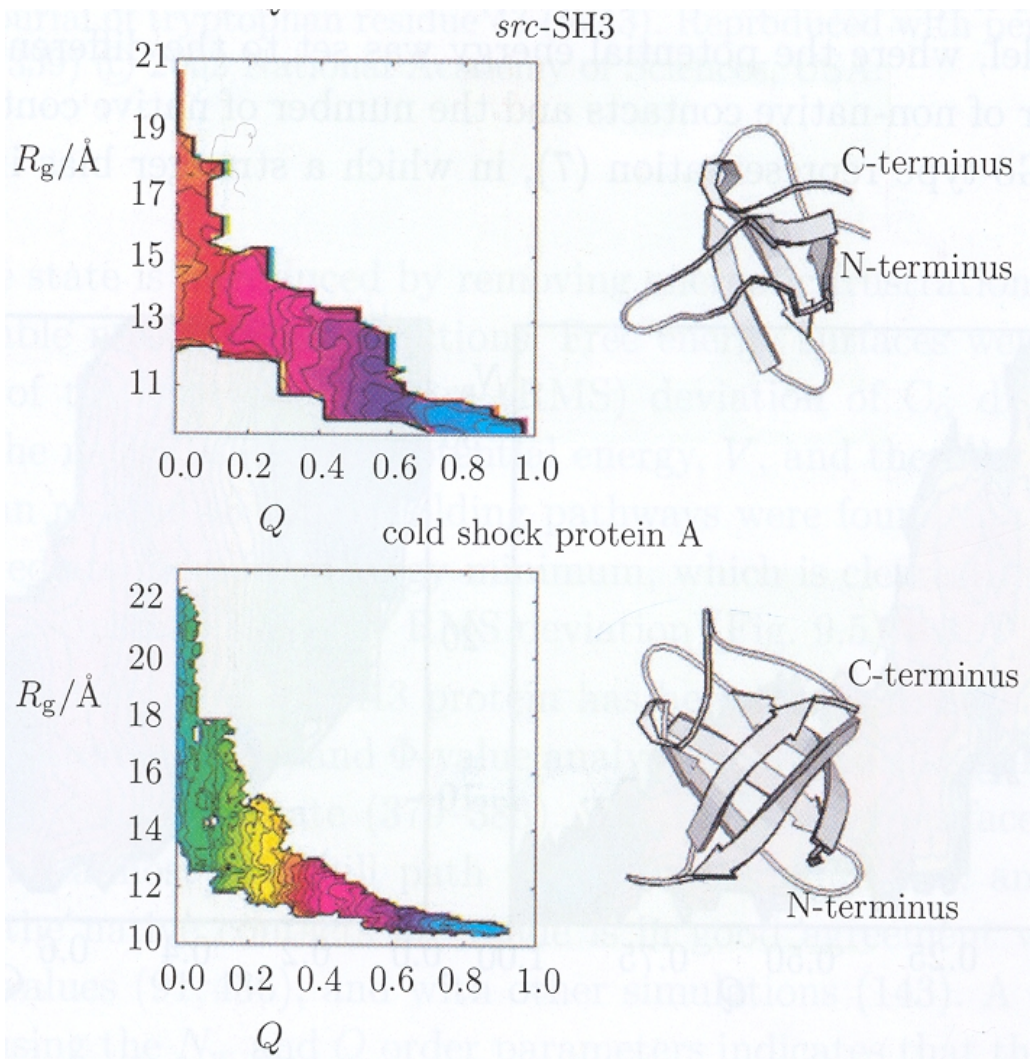
C.L.Brooks III,J. Am. Chem. Soc. 121, 9947 (1999)

...yet in other cases projection of trajectories on the two coordinates (radius of gyration),(number of native contacts) gives a reasonable idea of the folding process...

(CHARMM potential. C.L.Brooks III et al, Ann. Rev. Phys. Chem. 52, 499 (2001))

this:still bad





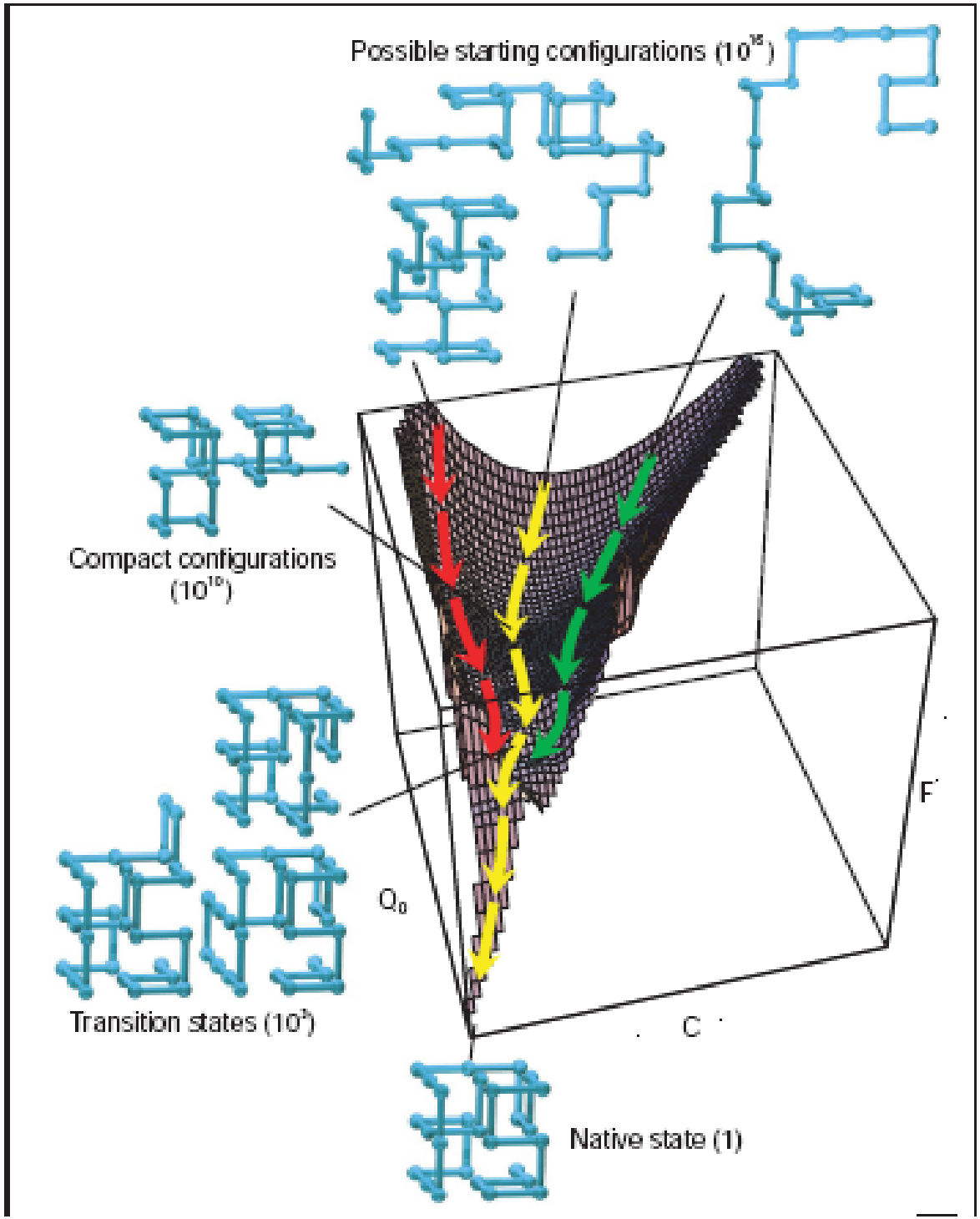
Using these coordinates allows a comparison between models of different nature.

Free-energy as a function of the number of native contacts (Q0) and the total number of contacts (C).

The native state is a cube with Q0 = 28 (100%).
Yellow= average path. Green and red trajectories include
~95% of the trajectories.

From 10^{16} possible random starting conformations, rapid collapse to a disordered globule; then slow, non-directed search among the 10^{10} semicompact conformations for one of approximately 10^{3} transition states.

A.R.Dinner et al, TIBS 2000



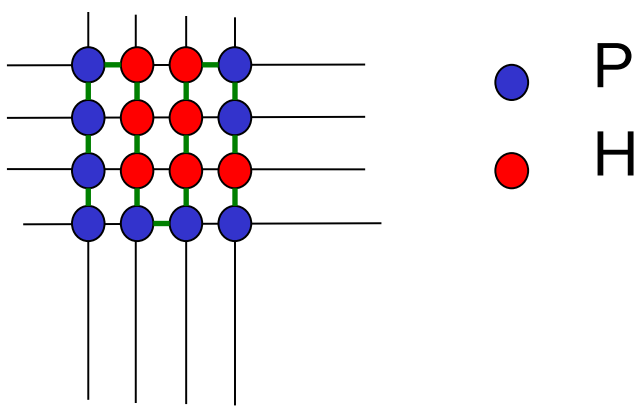
How shall we design sequences to obtain this kind of landscapes?

The HP model

- Two kinds of amino-acids: hydrophobic and polar (H and P).

- H amino-acids attract each other (it's an effective interaction!).
  $E_{HH}<E_{HP}<E_{PP}$ (actually, $E_{HH}+E_{PP}<2E_{HP}$, to favor segregation)

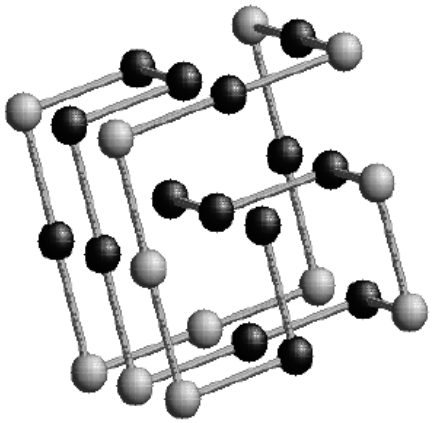
- Proteins are modeled on or off-lattice, in two or three dimensions

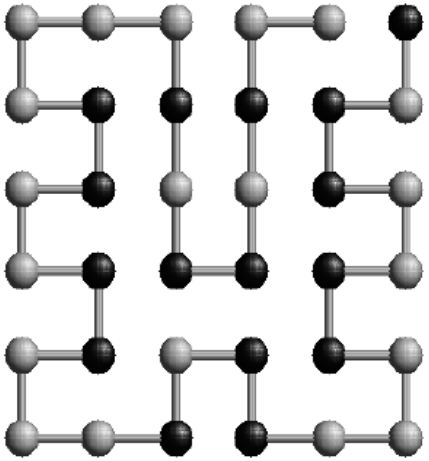


K. F. Lau and K. A. Dill, Proc. Natl. Acad. Sci. USA 87, 638 (1990),
H.Li et al, Science 273, 666 (1996)

Different structures show very different designabilities (=N_s of sequences having that structure as ground state)

(Results with $E_{HH} = -2.3$, $E_{HP} = -1$ and $E_{PP} = 0$
Below: most designable structures)





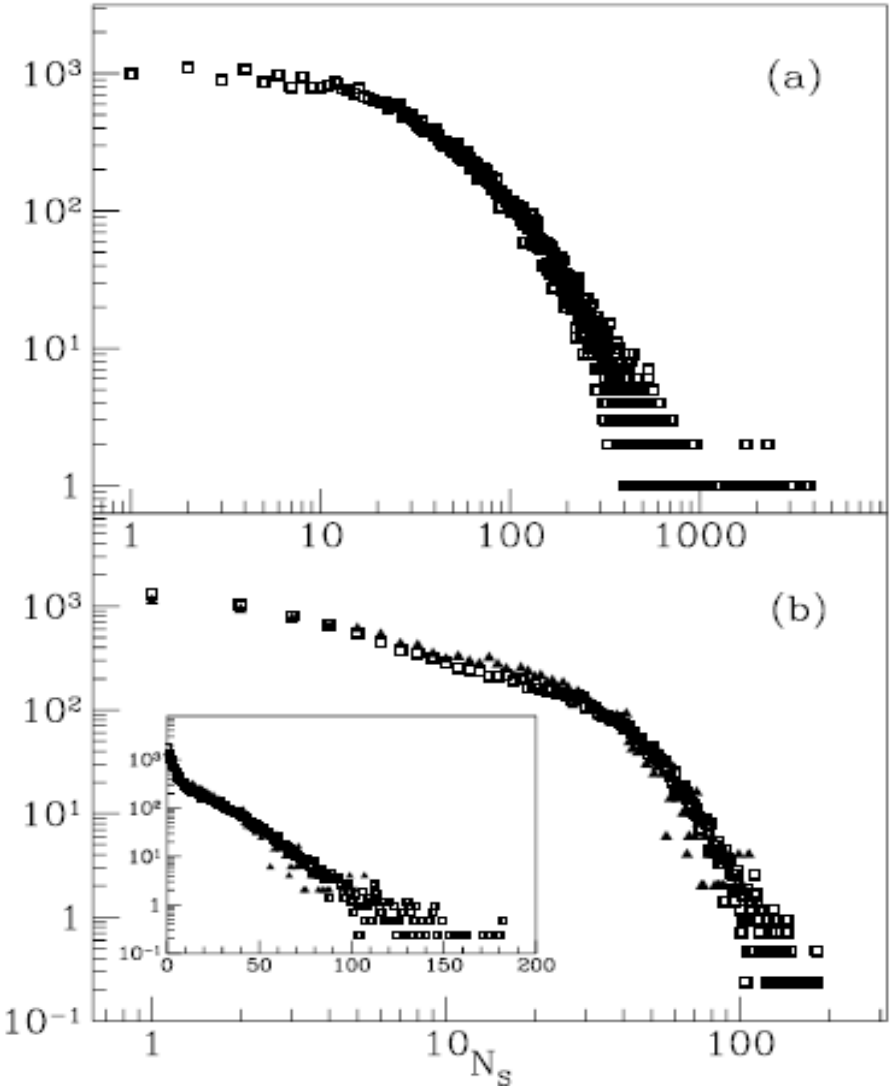
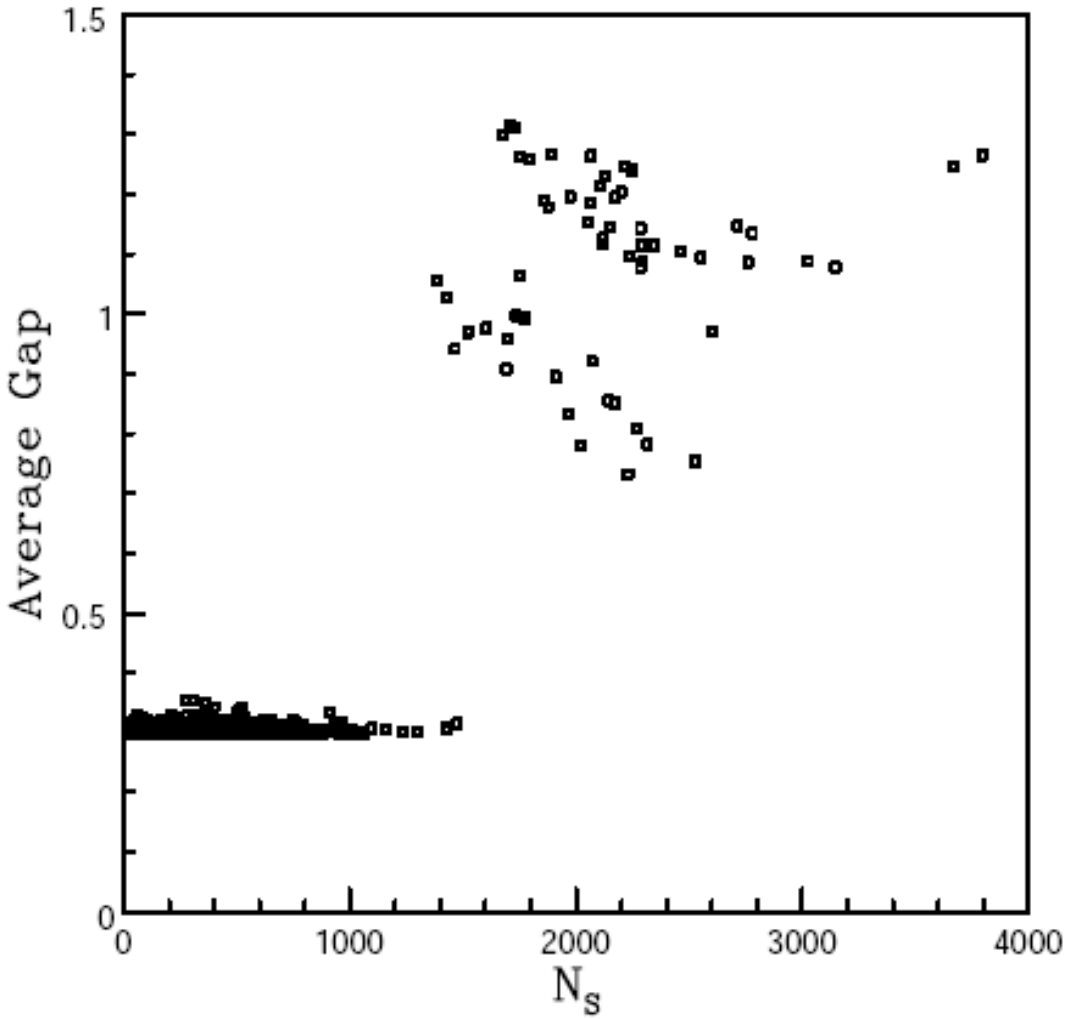


FIG. 1. (a): Histogram of number of structures with a given number of associated sequences N_S for 3D $3 \times 3 \times 3$ case, in a log-log plot. (b): Histogram of number of structures with a given N_S for 2D 6×5 (filled triangle) and 6×6 (open square) case, in a log-log plot. The bin size and rescaling of y axis are explained in the text. Insert: same data in a semi-log plot.

Most designable sequences are also the most thermodynamically stable, in the sense that they show the biggest gap between GS and the next low energy structure (which has unrelated geometry, due to lattice constraint).

Average gap of 3D 3x3x3 structures plotted against N_s of sequences folding to that structure.



Results:

The HP model predicts proteins with a hydrophobic core (as seen experimentally), that fold below a certain temperature.

Out of the 2N possible sequences (N is the protein length), just a few (perhaps too few...) are "good" sequences, with a unique native state (represented by the ground state).
So, we recover that only a small number of sequences can be proteins.

Designability:

Of all the possible compact conformations, only a few are chosen by the "good" sequences as native states. "Fold families" appears naturally within this model.

These results hold true even with Miyazawa-Jernigan interactions (20 letters), and in simple off-lattice models, but with discrete variables (Wingreen et al, Polymer 2004)

A simple, off-lattice model: BLN

D.Thirumalai et al, PNAS, 87, 3526 (1990), D.J.Wales, ”Energy Landscapes”, Cambridge Univ. Press,2003)

B=hydrophobic
L=hydrophilic
N=neutral

$$V = \frac{1}{2}K_r \sum_{i=1}^{N-1} (R_{i,i+1} - R_e)^2 + \frac{1}{2}K_\theta \sum_{i}^{N-2} (\theta_i - \theta_e)^2$$

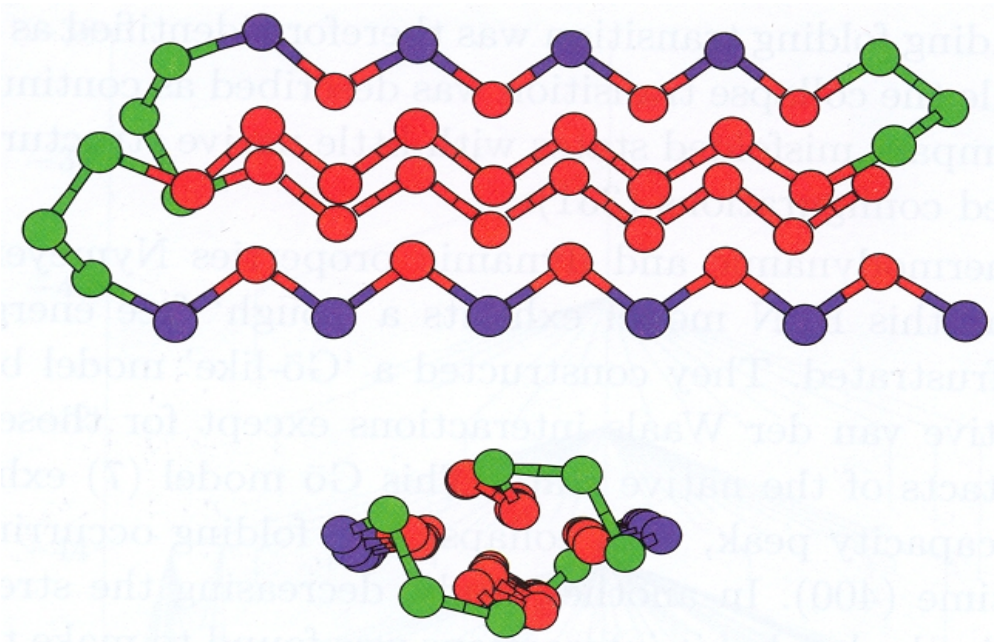
$$+ \epsilon \sum_{i}^{N-3} [A_i(1 + \cos\varphi_i) + B_i(1 + \cos 3\varphi_i)]$$

$$+ 4\epsilon \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} C_{ij} \left[\left(\frac{\sigma}{R_{ij}}\right)^{12} - D_{ij} \left(\frac{\sigma}{R_{ij}}\right)^{6} \right],$$

$$i, j \in \mathrm{B}, \quad C_{ij} = 1, \quad D_{ij} = 1,$$

$$i \in \mathrm{L},\ j \in \mathrm{L, B}, \quad C_{ij} = \frac{2}{3}, \quad D_{ij} = -1,$$

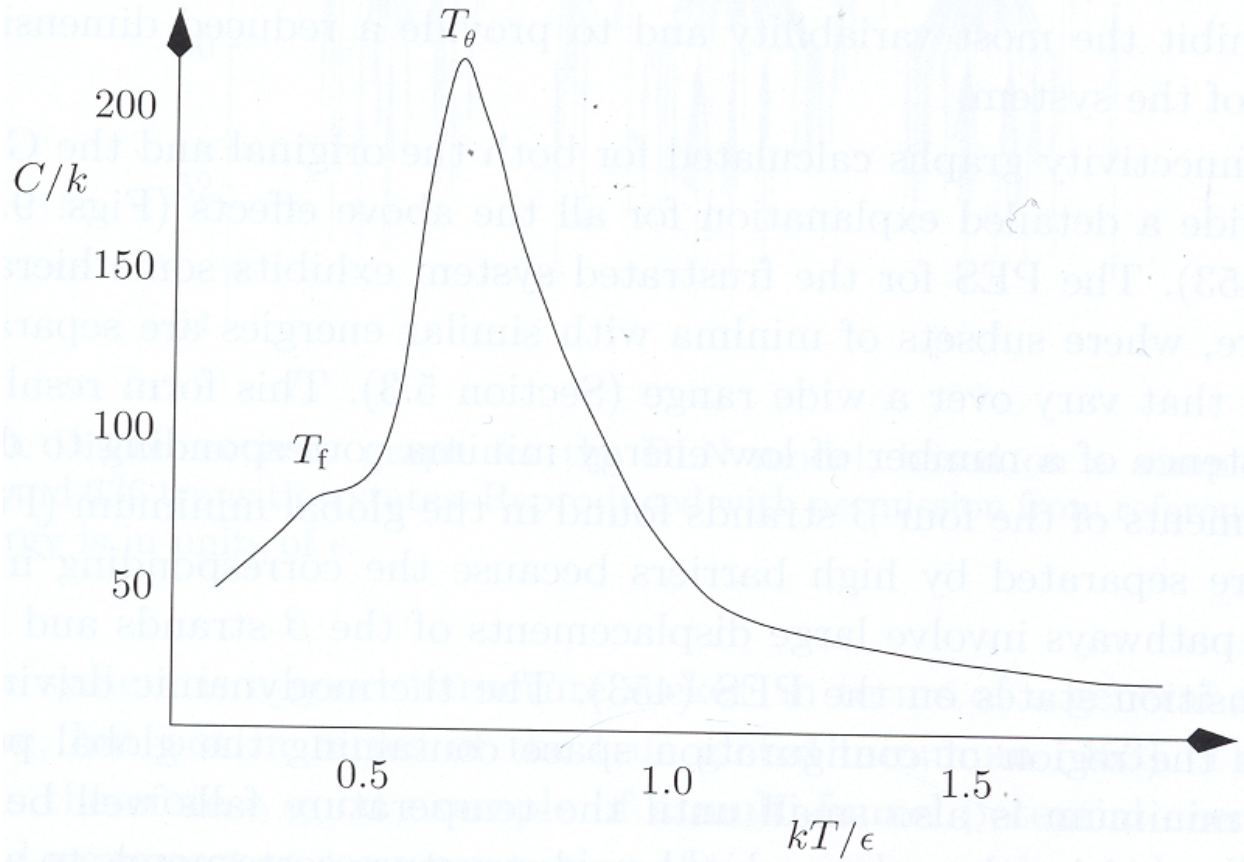
$$i \in \mathrm{N},\ j \in \mathrm{L, B, N}, \quad C_{ij} = 1, \quad D_{ij} = 0,$$



The sequence was designed to fold on the above structure

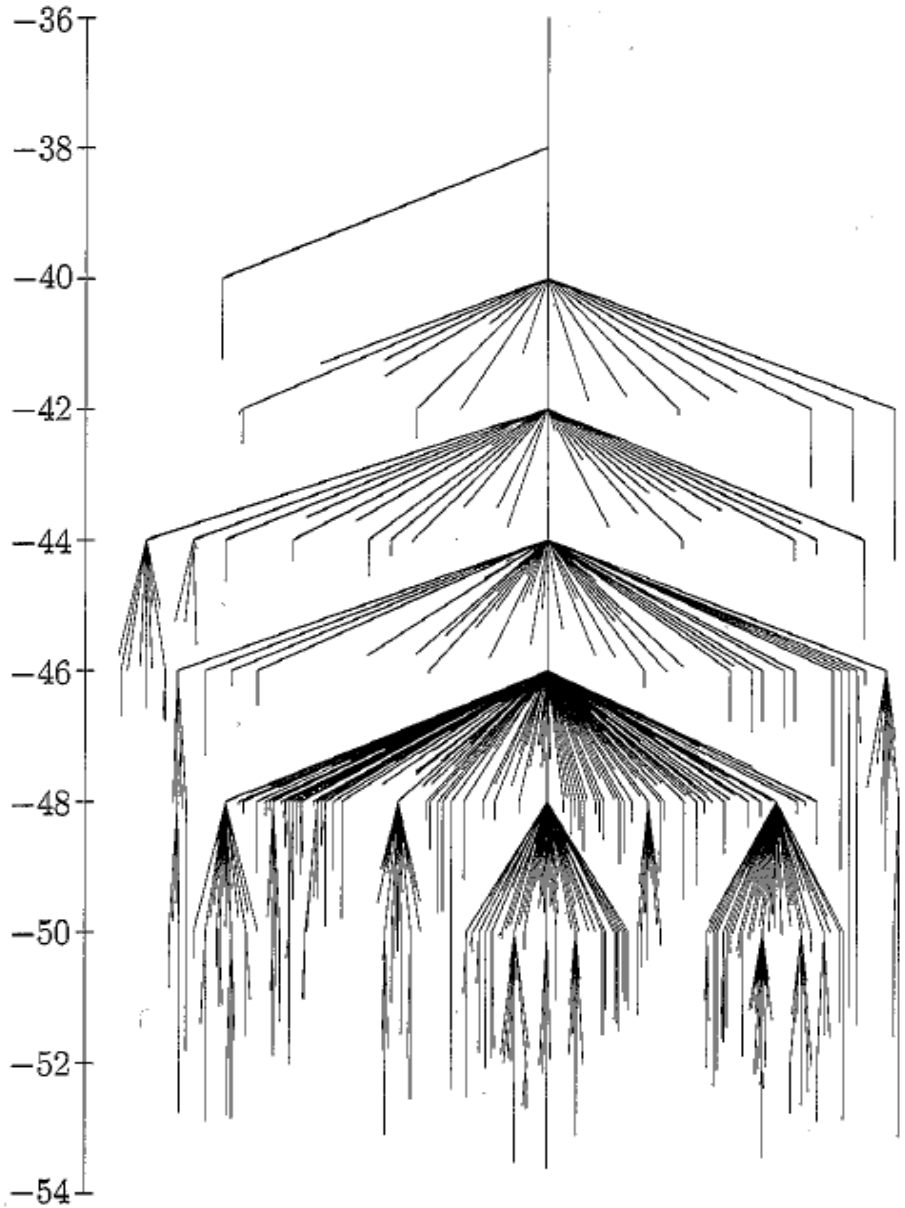
That is: attraction between BB, repulsion otherwise

Collapse transition at T_θ, followed by a folding one at T_f. The free-energy profile as a function of the potential energy reveals two minima at T_f: first-order like.

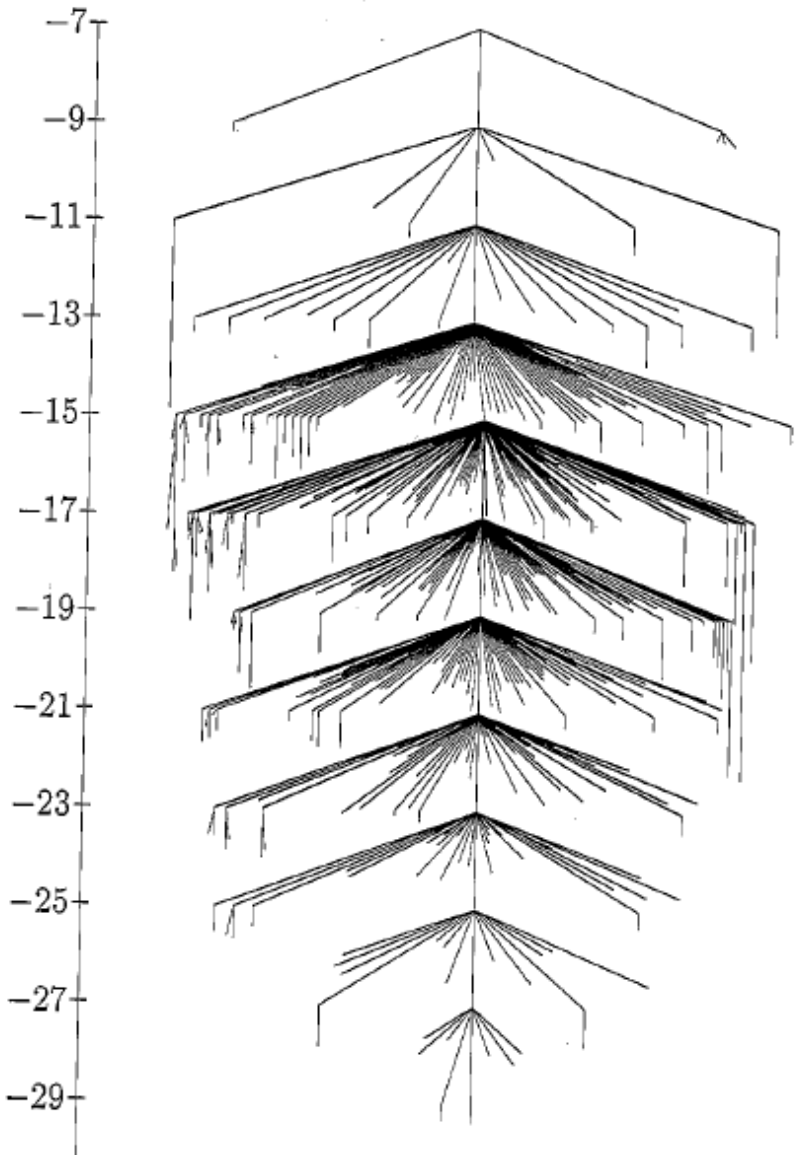


With a fine tuning of the interactions, it is possible to create a "protein-like" peptide, even if it is not an extremely good folder: different arrangements of the strands produce competing low-energy structures

BLN,500 minima, 636 TS:
Many competing low-energy basins



BLN-GO model, 500 minima,
805 TS. Funneled landscape



We have seen that we can take a reasonable structure, a small alphabet and can design a “reasonable” potential to fold that sequence into the target structure. Yet, most of the times we will have to face the problem of alternative minima for our sequence, givin rise to frustration and slow interconversion.

Let's take an alternative approach:
use a unreasonable potential, to guarantee that there is no energetic frustration (if present, frustration will be of just an entropic origin), and study the folding process in this “smoothed out” landscape...

Gō-like models:

from the knowledge of the native structure, build an hamiltonian which favors the native contacts and disregard all the others
(...are you kidding? No!)

These models assume the knowledge of native structure to address the following questions:
- Is energetic frustration really so relevant during folding? what happen if we delete it?
- How much do rates depend on just the native state geometry?
- Which geometrical characteristics are more relevant?
- Are there geometries which are more difficult to reach?

These models present a landscape as funneled as geometry allows!

This approach could be realistic if:
- real native contacts are better than non natives and/or
- during folding, dynamics only pass through good contacts

Is it realistic? Compare with experiments...

Go-like models (all atoms, reduced represent., Cα)

Example:
Cα case:

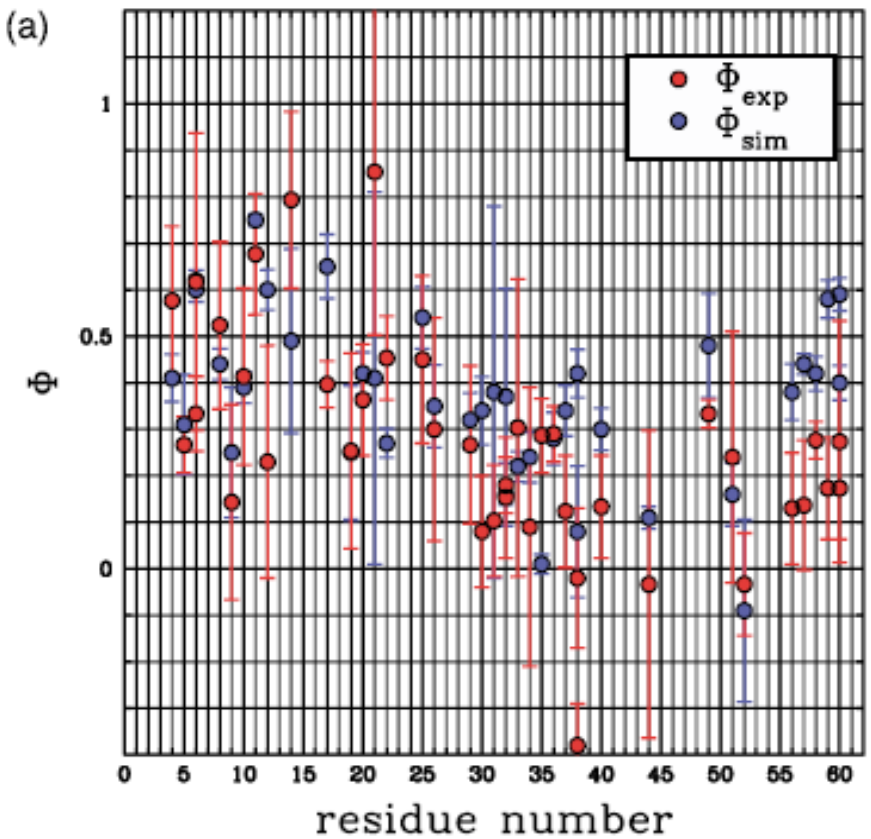
$$E(\Gamma,\Gamma_N)_{\text{Gō}} = \sum_{bonds} K_r (r-r_N)^2 + \sum_{angles} K_\theta (\theta-\theta_N)^2 + \sum_{dihedral} K_\phi^{(n)} [1+\cos(n\times(\phi-\phi_0))] + \sum_{i<j-3} \left\{ \epsilon_1(i,j) \left[6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} - 5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \right] + \epsilon_2(i,j)\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \right\}$$

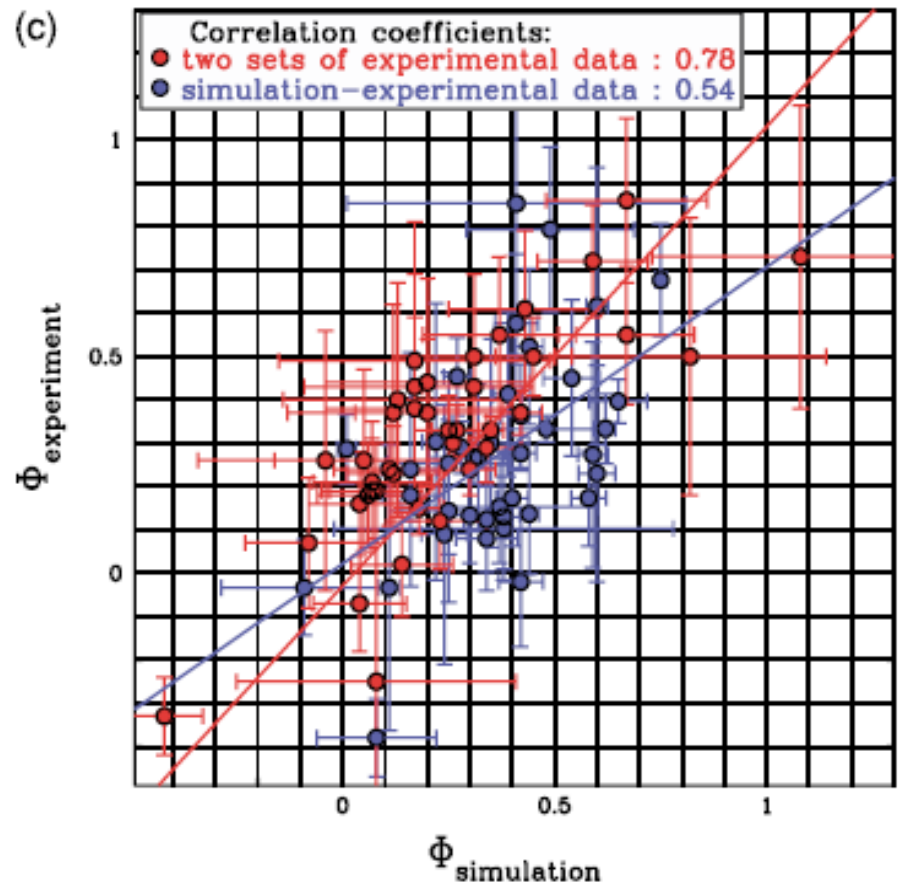
if i,j native pair: $\varepsilon_1(i,j) < 0$, $\varepsilon_2(i,j) = 0$, otherwise $\varepsilon_1(i,j) = 0$, $\varepsilon_2(i,j) > 0$

σ_{ij}=contact distance if i,j native pair,
otherwise σ_{ij}=4 Å

With these models, it is possible to study folding routes, intermediates, ϕ-values of a protein ...

Clementi et al, J. Mol. Biol. 326, 933 (2003)





Introducing non-native interactions as a small perturbation:

(Clementi, Plotkin, Prot. Sci. 13,1750 (2004))

$$E(\Gamma,\Gamma_N)_b = E(\Gamma,\Gamma_N)_{G\bar{o}} + \sum_{non-native(i,j)} V_{nn}(r_{i,j},\{\eta_b\})$$

$$V_{nn}(r_{ij}) = \begin{cases} \left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} + \eta_{i,j}\left[1-\frac{1}{2}\left(\frac{r_{i,j}}{r_N}\right)^{20}\right] & \text{if } r_{i,j} < r_N, \\ \left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} + \frac{\eta_{i,j}}{2}\left(\frac{r_N}{r_{i,j}}\right)^{20} & \text{if } r_{i,j} > r_N. \end{cases}$$

η_{ij} = random gaussian variable with mean ε_{NN} and variance b^2.

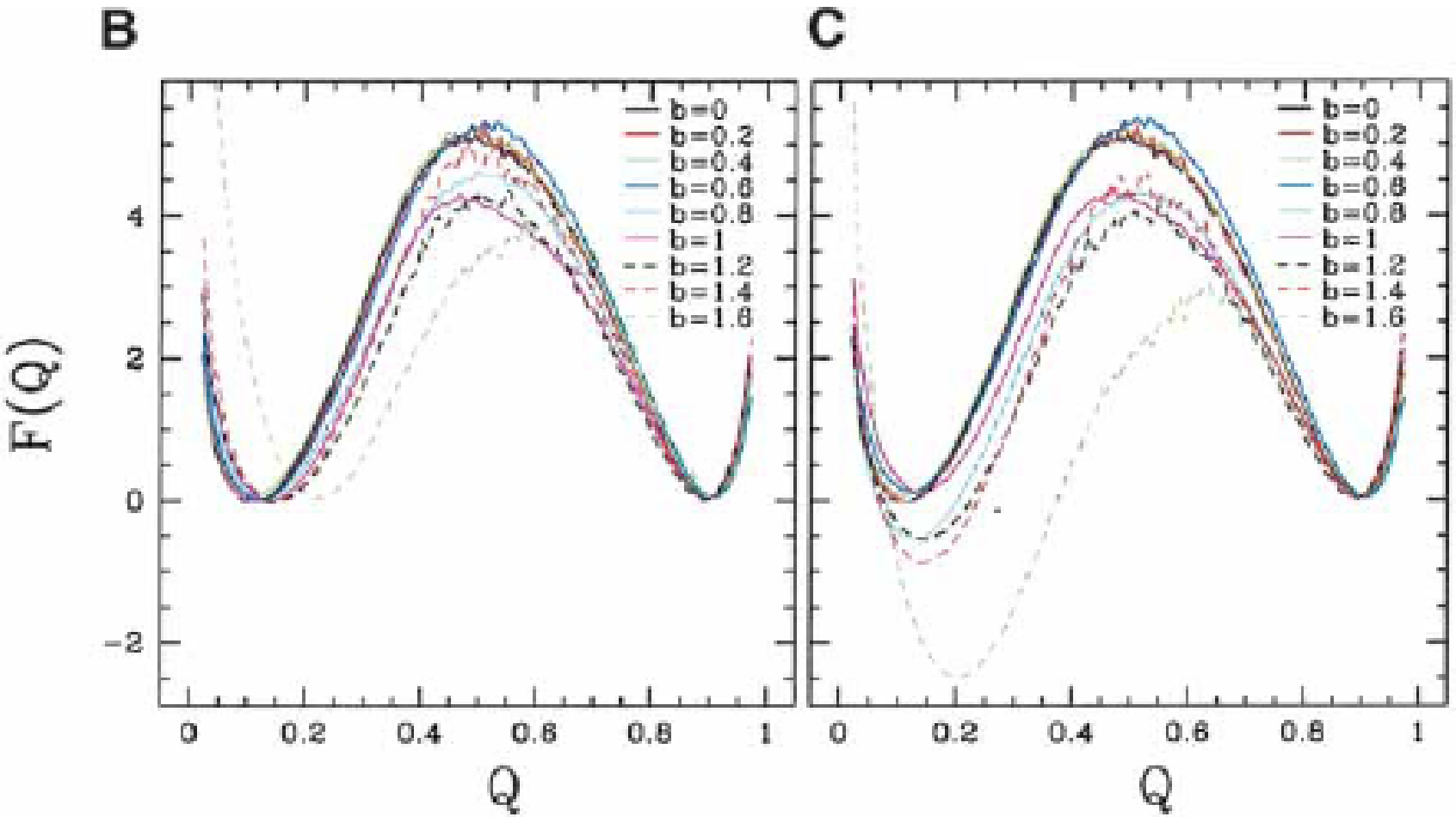
Choose: $\varepsilon_{NN} << \varepsilon$ to ensure $T_\theta \sim T_f$

$b << \varepsilon$ to ensure $T_g << T_f$

with ε the average energy per residue in the Go-model

B) barriers at T_m tend to decrease !

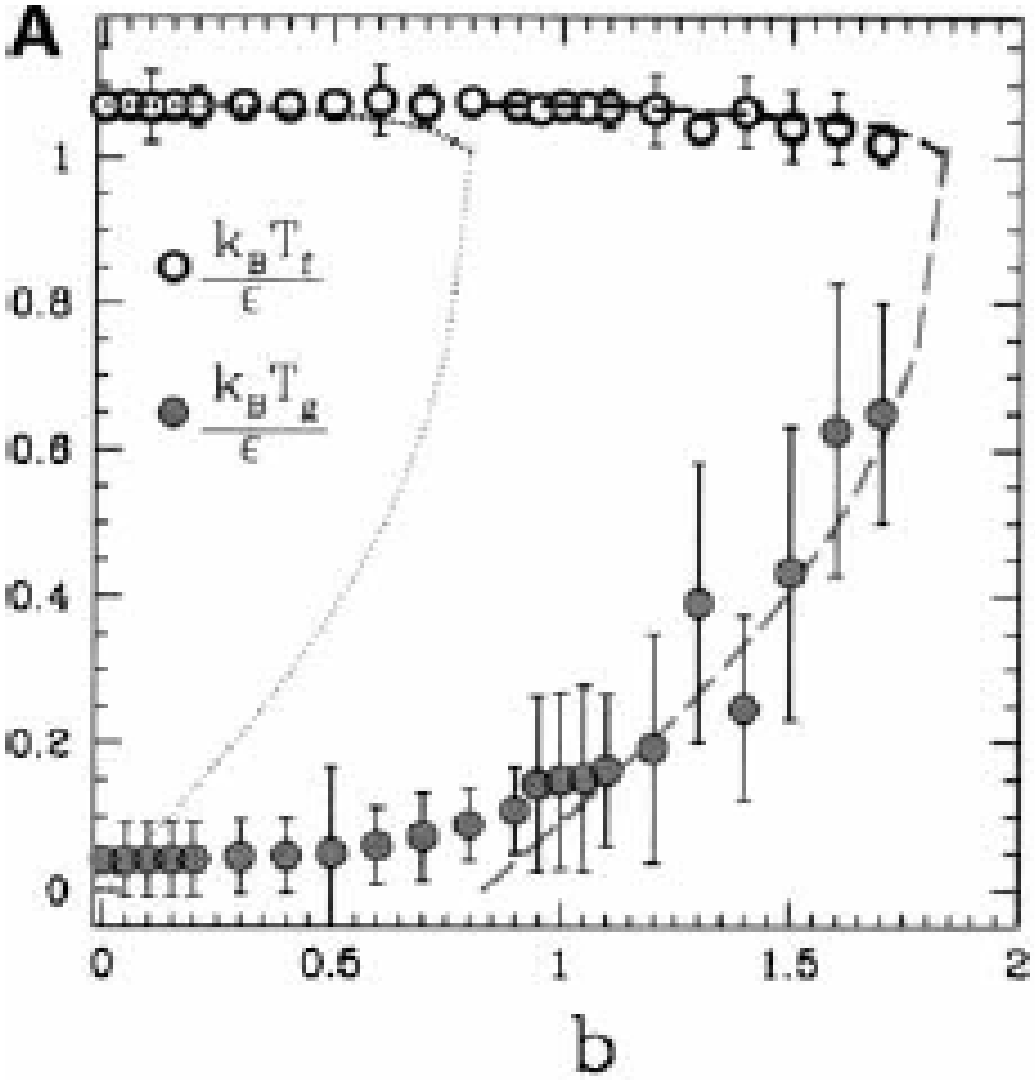
C) but at fixed $T=T_m(Go)$, we see that increasing variance of non-native contacts favors unfolding



Q=fraction of native contacts formed.

Results of simulations with Src-SH3 domain

...as seen with REM, in a good folder Tf/Tg>>1, and, as expected, the temperature of glass transition T_g increases with b!



Results:

- with good tuning of parameters, these models can recover NS

- substantially correct identification of experimental folding nucleus and behavior (which structure forms first, etc.) in a number of proteins (CI2, SH3, Barnase, circular permutations...)

- with all atoms, improved cooperativity

Small proteins appear indeed to have little energetic frustration!

Adding moderate non-native interactions, folding is faster, provided that the folding temperature significantly greater than the glass one.

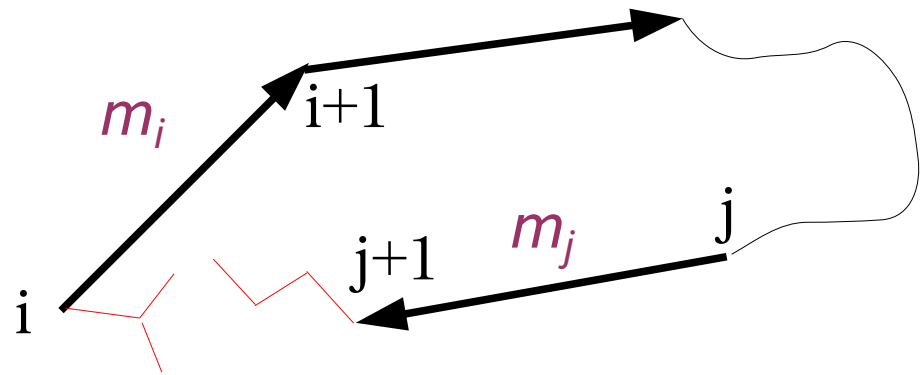
Discrete Go-models: Wako-Saitô-Muñoz-Eaton (WSME) Model

H. Wako and N. Saito, J. Phys. Soc. Jpn 44, 1931 (1978);
V. Muñoz et al, Nature 390, 196 (1997).

$$H_{eff}(\{m_k\}) = \sum_{i<j=2}^{N} \varepsilon \Delta_{ij} \prod_{i \leqslant k \leqslant j} m_k + RT \sum_{i=1}^{N} q_i m_i$$

m_i = 0,1 (peptide unit between residue i, i+1 is unfolded/native)

Δ_{ij} = number of atoms from residues i, j+1 closer than 4 Å in the native structure



ε = energy from each atom-atom contact (ε<0)

q_i = entropic cost of native bond m_i =1 with respect to m_i =0.

$\prod_{i \leqslant k \leqslant j} m_k$ ensures that only contacts within a native stretch are accounted

Equilibrium solution: define $x_{i,j} = \prod_{i \leqslant k \leqslant j} m_k$

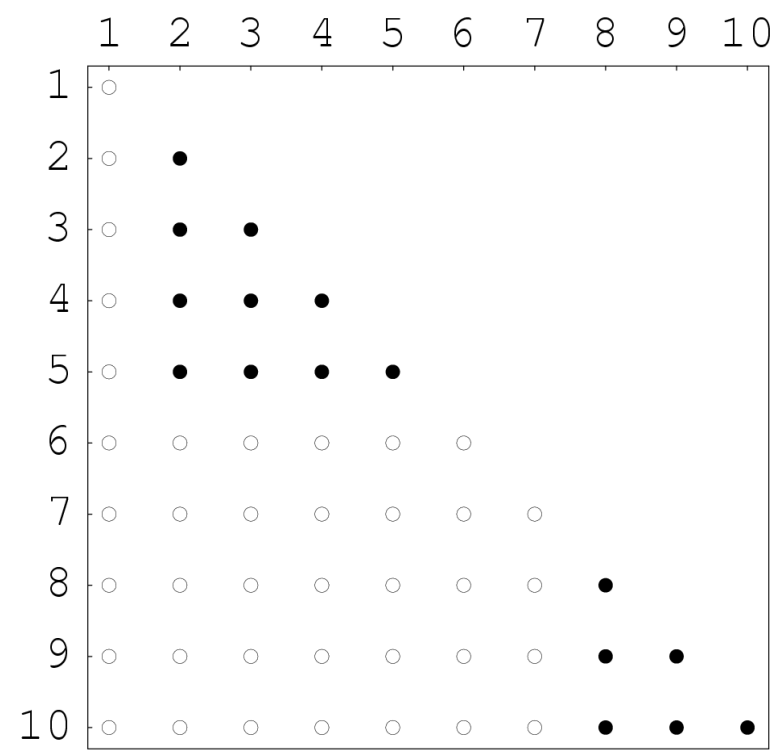
$$H_{eff}(\{m_k\}) = \sum_{i<j=2}^{N} \varepsilon \Delta_{ij} x_{i,j} + RT \sum_{i=1}^{N} q_i x_{i,i} = \sum_{i \leqslant j} h_{i,j} x_{i,j}$$

apparently uncoupled, but:

Constraint:

$$x_{i,j} = x_{i+1,j} \; x_{i,j-1}$$

This enforce the "black triangle" shape on the right:



$\bullet : x_{i,j} = 1 \qquad \circ : x_{i,j} = 0$

Equilibrium solution (continued):

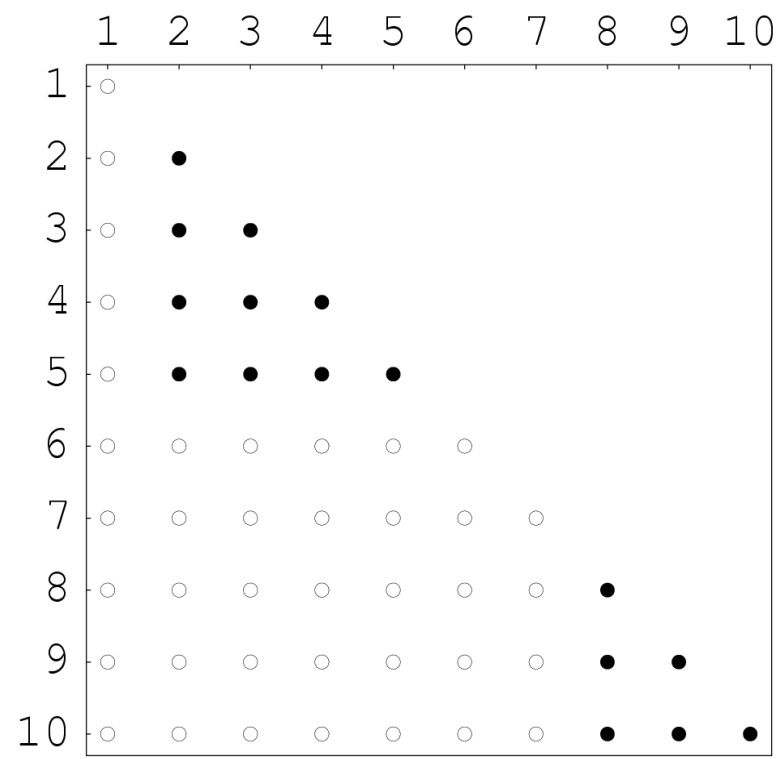
Transfer matrix approach: (P.Bruscolini, A.Pelizzola, PRL 2002)

- row j (j=1...N) can exist in a total of j+1 states :

$\mu_j = 0, \ldots,$ *j*right-aligned black balls

We can write the partition function:

$$Z = \sum_{\{\mu_k\}} Q^{(0,1)}_{0,\mu_1} \, Q^{(1,2)}_{\mu_1,\mu_2} \ldots Q^{(N-1,N)}_{\mu_{N-1},\mu_N} \, Q^{(N,N+1)}_{\mu_N,0}$$

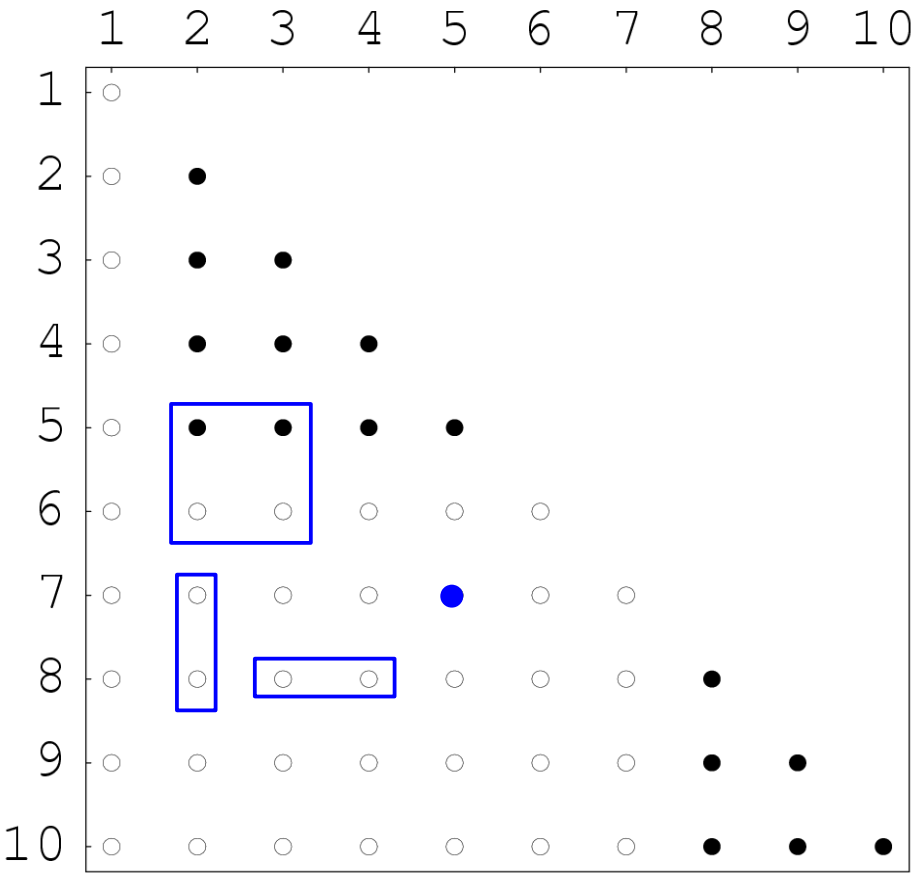


$\bullet : x_{i,j} = 1 \qquad \circ : x_{i,j} = 0$

Equilibrium solution (continued):

Cluster Variation Method :

$$p(\boldsymbol{x}) = \frac{\prod_{\square} p_{\square}(x_{\square}) \prod_{\bullet} p_{\bullet}(x_{\bullet})}{\prod_{\square} p_{\square}(x_{\square}) \prod_{\square} p_{\square}(x_{\square})}$$



Free energy is calculated minimizing:

$$F_{\mathrm{CVM}} = \sum_{\alpha = \square, \square, \square, \bullet} \sum_{x_\alpha} \left(p_\alpha(x_\alpha) h_\alpha + a_\alpha p_\alpha(x_\alpha) \ln p_\alpha(x_\alpha) \right)$$

Kinetics:

consider the discrete-time master equation

$$p_{t+1}(\boldsymbol{x}) = \sum_{\boldsymbol{x}'} W(\boldsymbol{x}' \to \boldsymbol{x}) p_t(\boldsymbol{x}')$$

where

- $\boldsymbol{x} = \{x_{i,j},\ 1 \leqslant i \leqslant j \leqslant N\}$ is any allowed configuration of the $x_{i,j}$,
- $\boldsymbol{x}'$ differs from $\boldsymbol{x}$ by the folding flip of just one peptide unit,
- the transition probability W is given by Metropolis Kinetics:

$$W(\boldsymbol{x}' \to \boldsymbol{x}) = \tau^{-1} \min(1, \exp[-\beta(H(\boldsymbol{x}) - H(\boldsymbol{x}')]))$$

$$W(\boldsymbol{x} \to \boldsymbol{x}) = 1 - \sum_{\boldsymbol{x}' \neq \boldsymbol{x}} W(\boldsymbol{x} \to \boldsymbol{x}')$$

Two possible approaches:

1) Monte Carlo Simulations

2) Local Equilibrium Approximation

$$p_t(\boldsymbol{x}) = \frac{\prod_{\square} p_{t,\square}(x_{\square}) \prod_{\bullet} p_{t,\bullet}(x_{\bullet})}{\prod_{\square} p_{t,\square}(x_{\square}) \prod_{\square} p_{t,\square}(x_{\square})}$$

Imposing this factorization at all times, it is possible to write an evolution equation for the observables $\langle x_{ij} \rangle$:

$$\frac{d\langle x_{i,j}\rangle}{dt} = f(\{\langle x_{k,k}\rangle\})$$

(Pelizzola, Zamparo PRL. 2006; J. Stat. Mech. 2006)

Experimental rate dependence on the "contact order":

(Ivankov et al, Prot. Sci. 12, 2057 (2003))

$$ACO = \frac{\sum_{i<j} |j-i| \Delta_{ij}}{\sum_{i<j} \Delta_{ij}}$$

circles: proteins with two-state kinetics
triangles: multi-state
crosses:small peptides

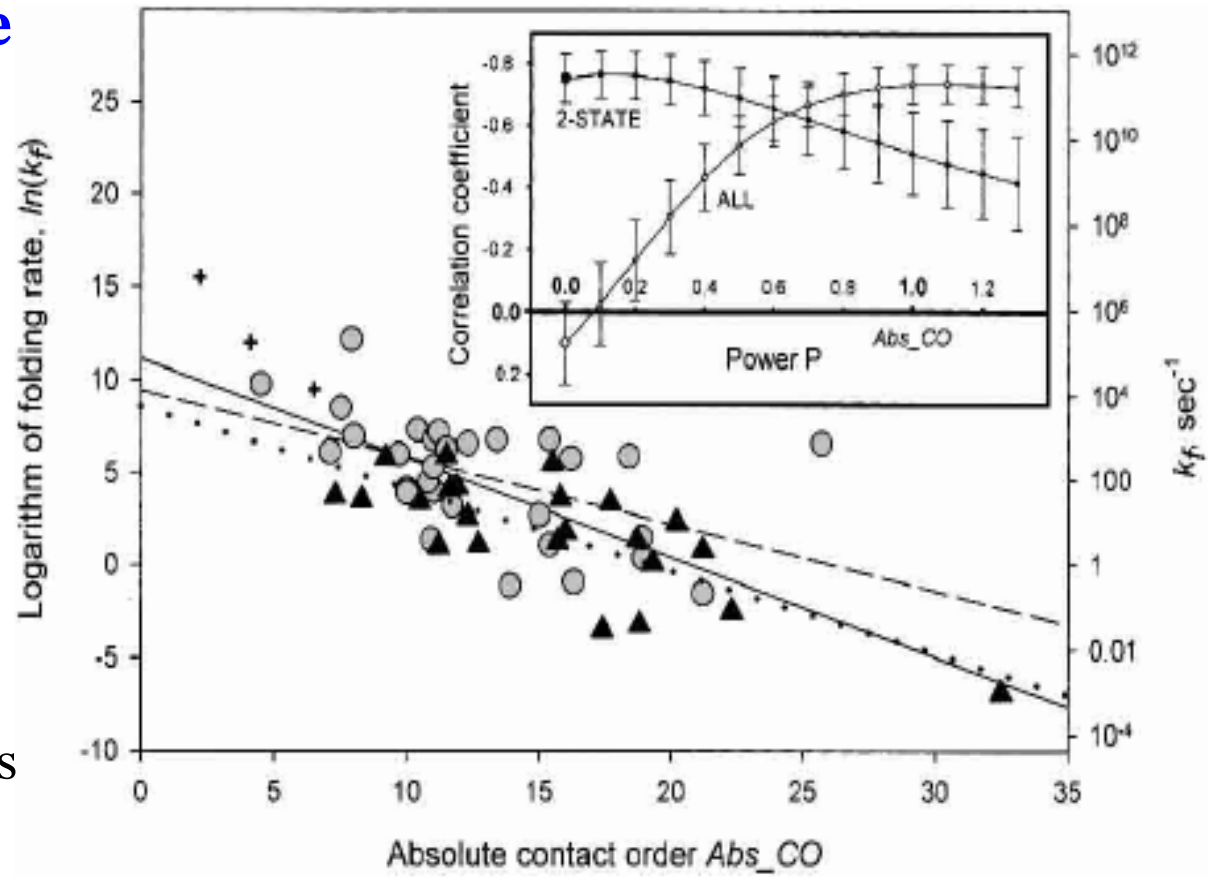
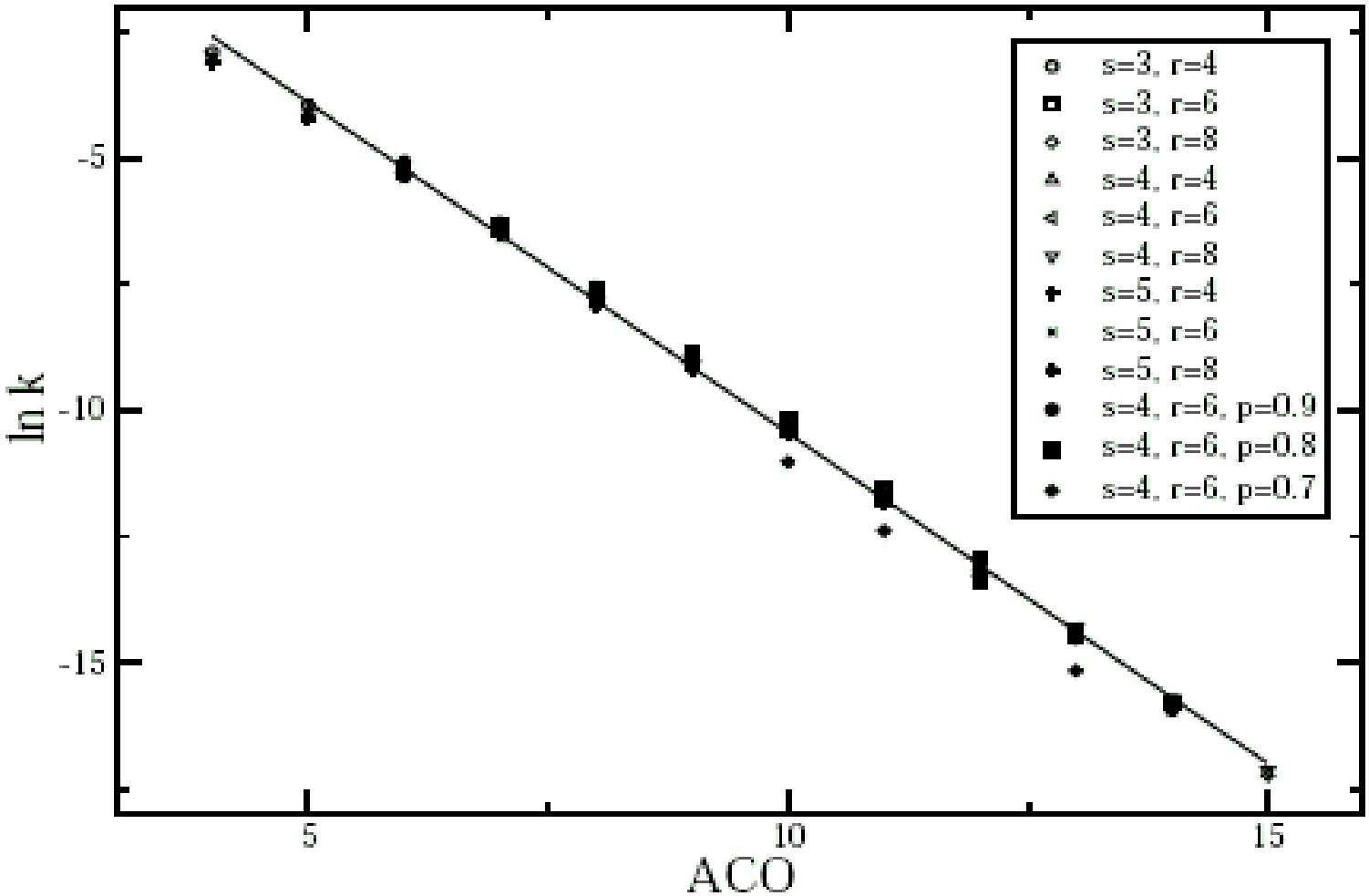


Figure 3. Logarithm of observed folding rate in water $ln(k_f)$ versus $Abs_CO = CO \times L$. See legend to Figure 1 for specification of the symbols and other details. The dashed line represents the best linear fit for two-state folders only (the fitted dependence is $y = 9.44 - 0.36x$; the correlation coefficient is –0.51); the dotted line represents the best linear fit for multistate folders only (the fitted dependence is $y = 8.56 - 0.44x$; the correlation coefficient is –0.78); the solid line represents the best linear fit for the totality of all peptides and proteins presented (the fitted dependence is $y = 11.15 - 0.54x$; the correlation coefficient is –0.74). (*Inset*) Correlation

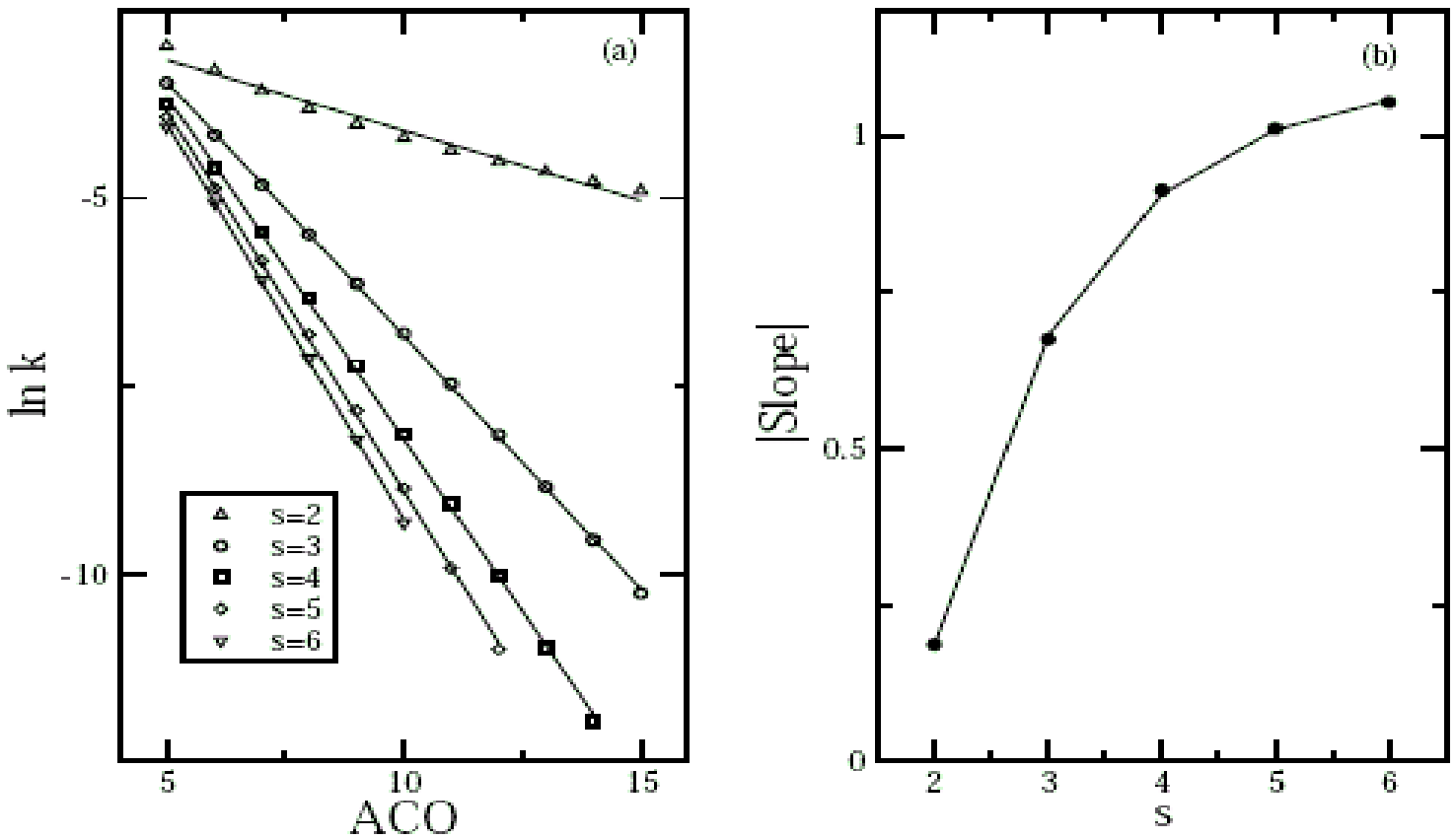
Elucidating the role of geometry on relaxation rates: ideal structures

(P.Bruscolini, A.Pelizzola, M.Zamparo, submitted)



$$\ln k = 2.6633 - 1.3113\ \mathrm{ACO}$$

Antiparallel structures: s= number of strands



$$|\text{Slope}| = 1.21 - 3.5 \exp(-0.62\, s)$$

Remarks

WSME model's thermodynamics can be known exactly, kinetics can be dealt with within a sound analytical approximation

These kind of models allow to ask generic questions on protein-like behavior, and also to predict, semi-quantitatively, the folding behavior of specific proteins:

- Reasonable estimate of the barrier height for simple proteins.

- Reasonable agreement with experimental results for identification of the folding nucleus (that is, the top of the barrier, that determines the speed of folding)

-Study of model structures, to understand the origin of specific features of real proteins

Conclusions

Proteins have evolved to cope with a number of requirements:
fast folding, stable but not too much, robust against mutations.
Simple macroscopic behavior from an amazing microscopic complexity.

Structure is such that energetic frustration is minimized, and the folding process amounts to a search of a few good pathways.

Many interesting aspects: we have focused on general features of protein-like behavior, and on the coarse-grained models describing them.

Folding, and functioning, are possible thanks to a "design" of the landscape.
Proteins: a tiny subset of "simple" systems, within a huge sea of "complex" random polymers. Is it possible to define "protein-likeness", in terms of the characteristics of the configuration space? To define an equivalence relation between sequences folding to the same structure?
To find similar simple behavior in subset of other glassy systems?
...a comprehensive theory is still waiting to be developed....