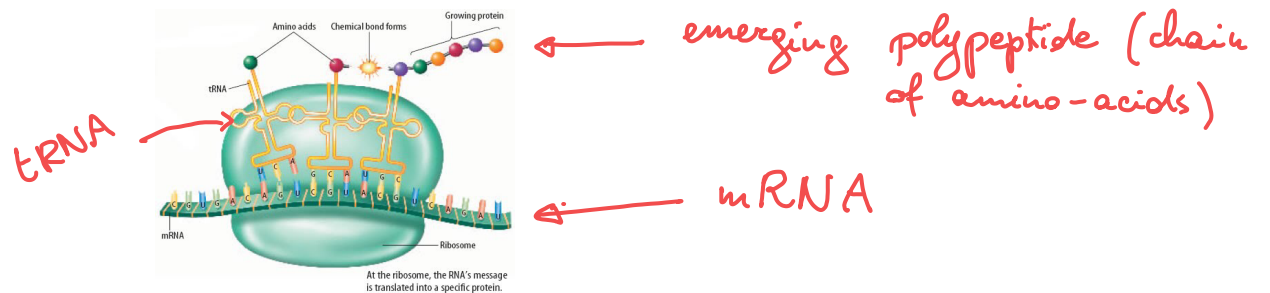


9. Protein Folding.

Proteins are produced by ribosomes, that read mRNA and translate it into a chain of amino-acids.



This chain comes out of the ribosome in an unfolded conformation, similar to a generic random coil. From there it must collapse (see chapter 8) but in a very specific conformation called the "native state".

We have seen that collapse is due generically to attractive forces. How can we make it unique?

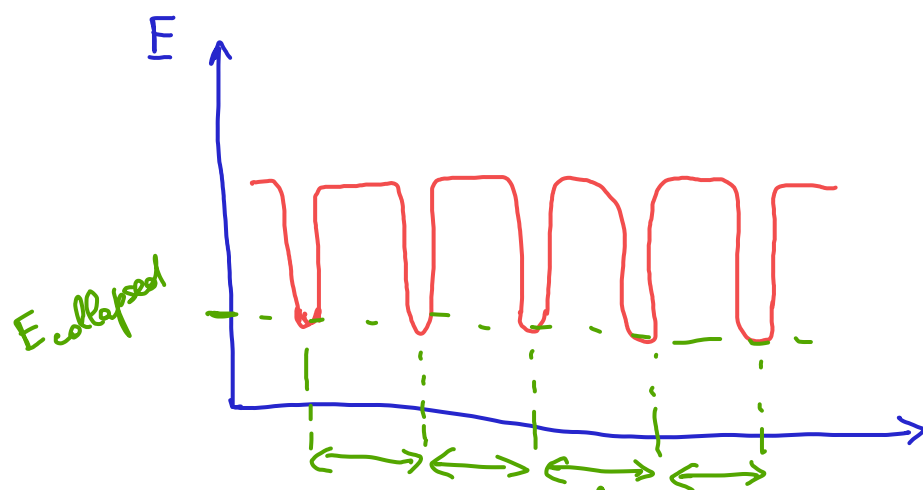
Question 1 : how many compact conformations are accessible to a homopolymer with attractive interactions between monomers?

It grows exponentially!

N monomers $\Rightarrow \mu^N$ compact conformations
($\mu > 1$; $\mu \approx 2$)

Energetically they are all equivalent, because if they are equally compact they have the same number of contacts per monomer.

We can think about the energy landscape of a homopolymer with attractive interactions as in this picture:



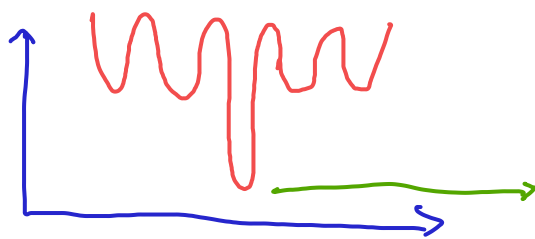
This distance represents the "number of moves" necessary to change one compact state into another:



They cannot be turned into each other by minimal moves.

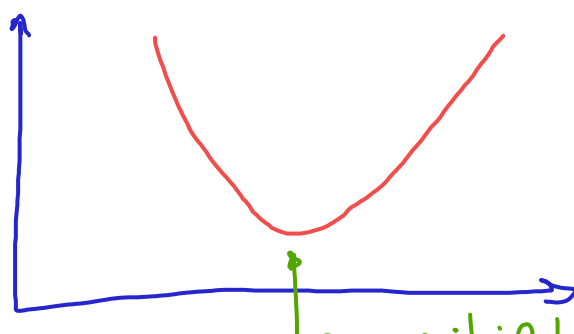
In order to produce a protein that collapses into a single state we need

i)



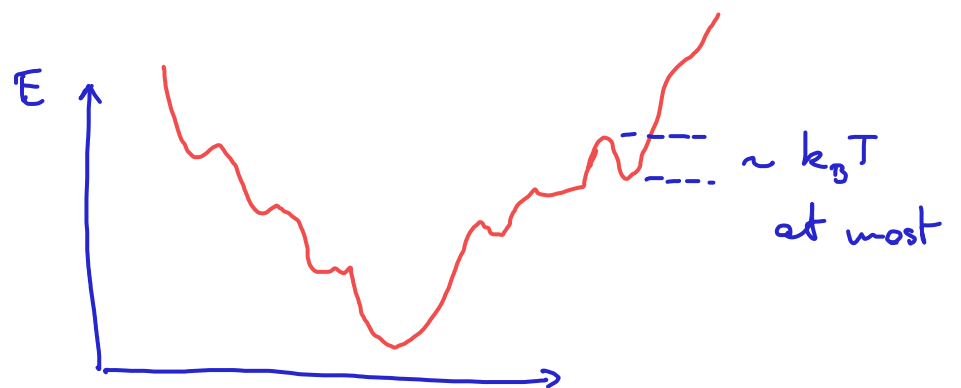
native state, single absolute minimum, necessary for thermodynamic stability

ii)



accessibility: no traps along the pathway

Something in between:



This brings to the next question:

Question 2

How can we sculpt the energy landscape?

We must first understand what are the building-blocks that we can use: the amino-acids.

All proteins in all organisms are made of 20 different kinds of amino-acid

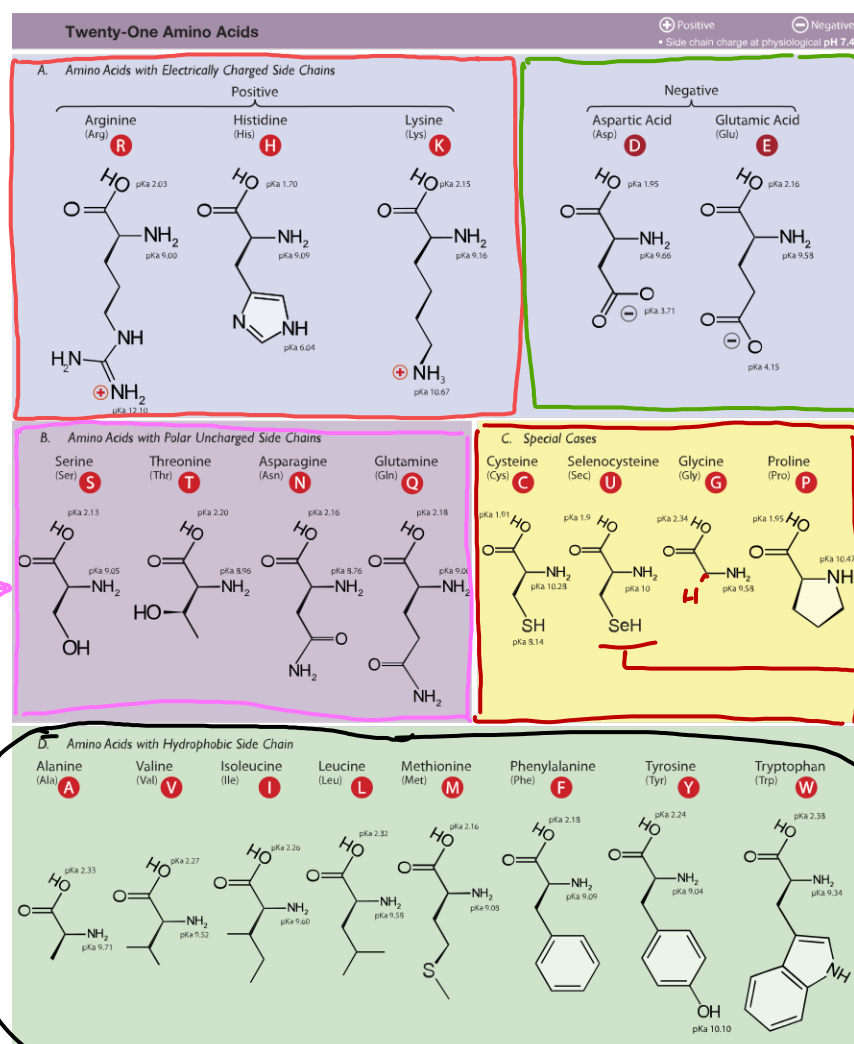
positively charged
they grab a proton from water

negatively charged:
they lose a proton to water

not charged,
but with
+ and - different
centers of mass
→ polar

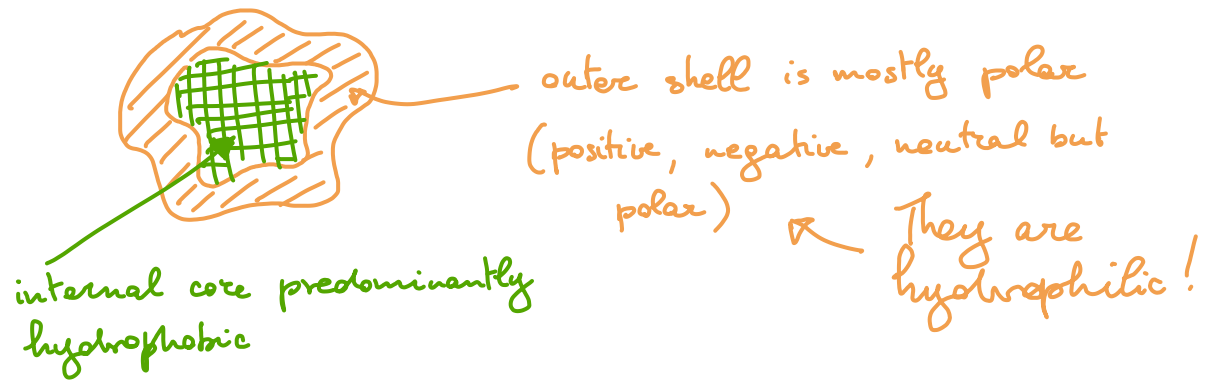
These are special
cases

is the 21st amino
acid, used only
by very few bacteria



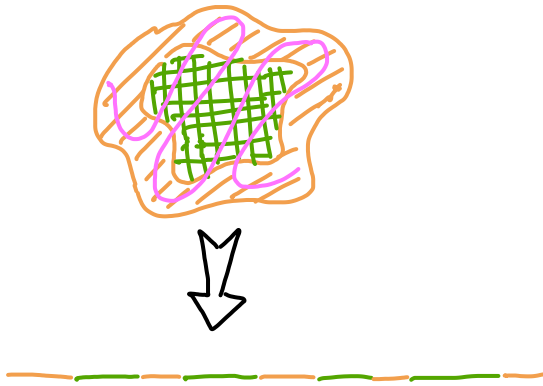
These amino-acids are non-polar ⇒ hydrophobic!

What drives the collapse of proteins?

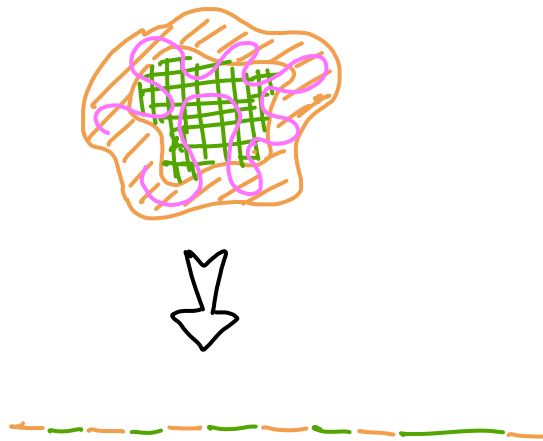


We can argue that the main driving force of folding is hydrophobicity.

The graph above neglects that this is actually a polymer.
Let's try "unfolding it"



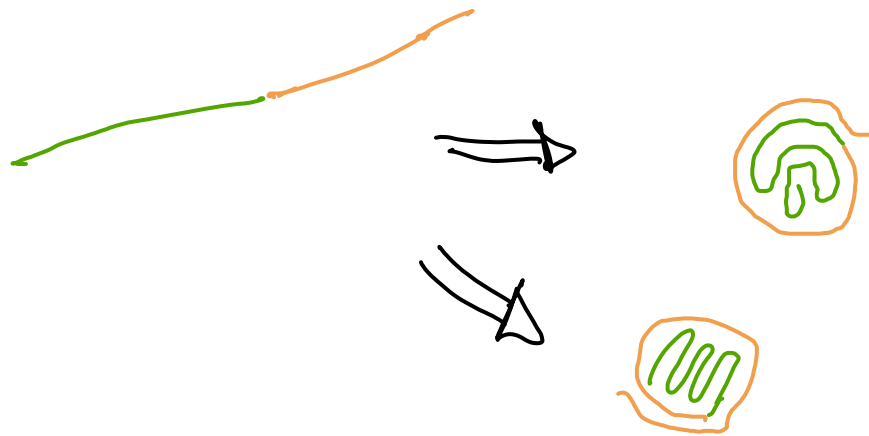
But also:



The different ways we can "sew" the chain require different arrangements of the amino-acids along the chain $\Rightarrow$ The sequence is crucial

Some sequences are not good as proteins:

e.g. the hydrophobic homopolymer
but also a co-polymer, first half of a kind (e.g. hydrophobic), and the second of the other (e.g. hydrophilic)

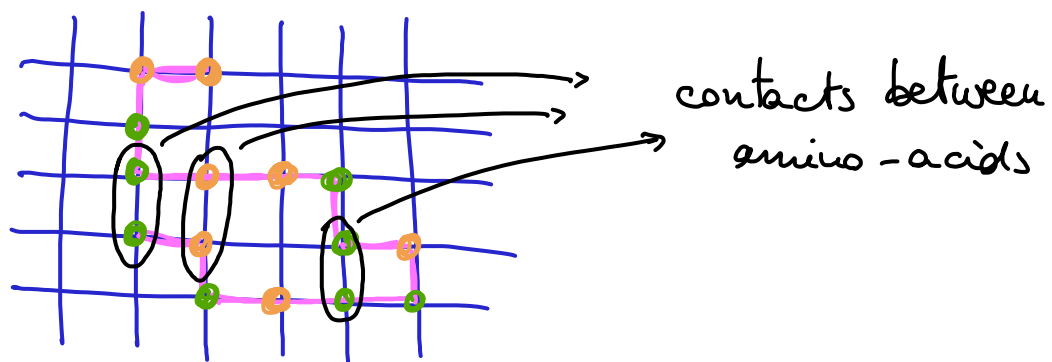


The hydrophobic core can be compactified in many different ways.

Essentially, only specific ways to alternate hydrophobic and hydrophilic amino-acids is going to produce polypeptides that fold on single, non-degenerate compact structures.

The first model trying to address this problem has been the HP model (H = hydrophobic, P = polar) proposed by Lau and Dill in 1989.

The 20 amino-acids are reduced to 2, and the polymers are drawn on a 2D lattice:



Each contact contributes to the energy :

$$\epsilon_{HH} = -\epsilon \quad (\epsilon > 0)$$

$$\epsilon_{HP} = -\alpha\epsilon$$

$$\epsilon_{PP} = 0 \quad \leftarrow \text{can always choose a reference energy.}$$

In general we can thus write the energy of a given sequence $\vec{S}$ ($\vec{S} = \{\sigma_1, \dots, \sigma_N\}$, for N monomers, with $\sigma_i = H$ or P), in conformation Γ , as

$$E(\Gamma, \vec{S}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N A_{ij} \epsilon_{\sigma_i \sigma_j}$$

$N = \# \text{ of monomers}$

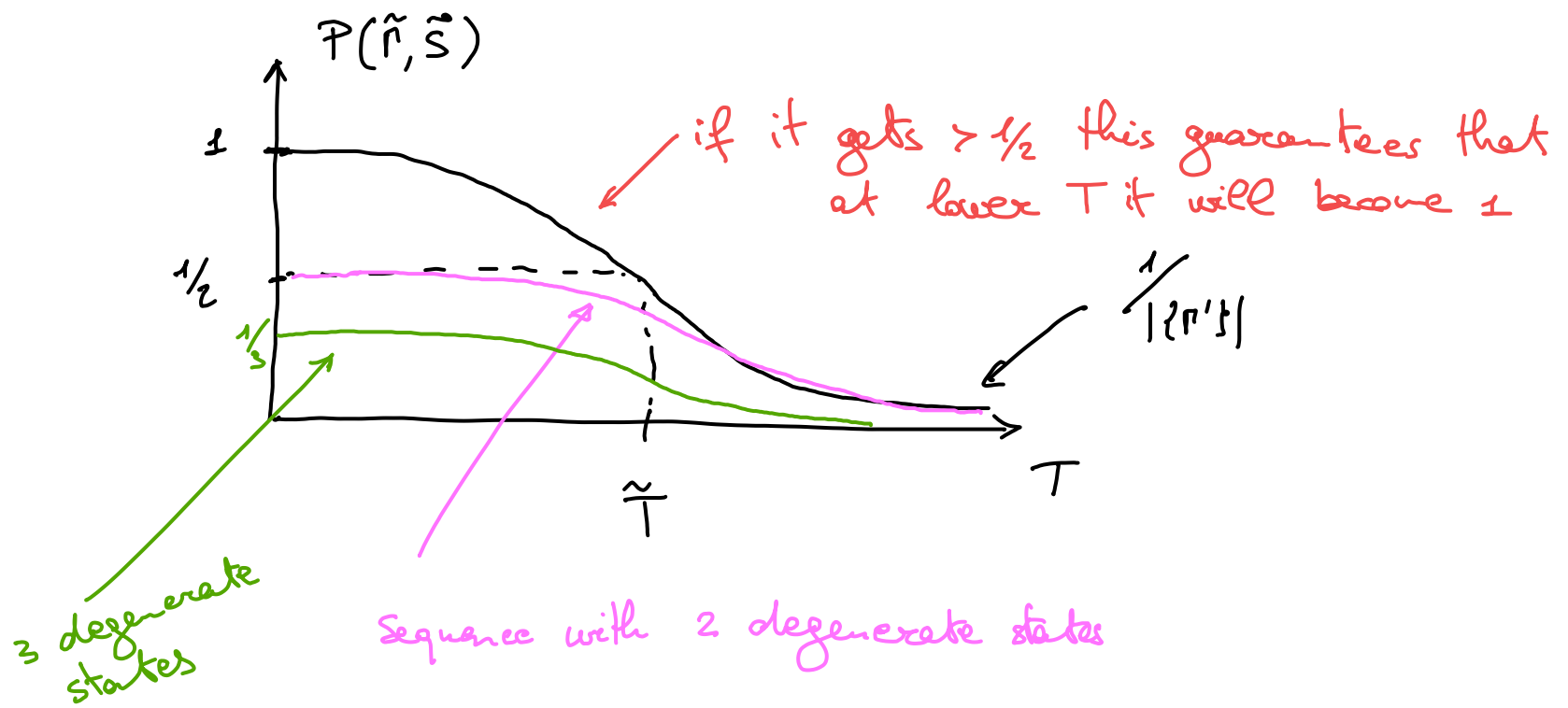
where A_{ij} is the contact matrix, telling whether amino-acids (monomers) i and j are nearest neighbors on the lattice (but not consecutive) or not :

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ n.n. of } j, \text{ but not consecutive} \\ 0 & \text{otherwise} \end{cases}$$

The probability of a given conformation is thus

$$P(\Gamma, \vec{S}) = \frac{e^{-\beta E(\Gamma, \vec{S})}}{\sum_{\{\Gamma'\}} e^{-\beta E(\Gamma', \vec{S})}}$$

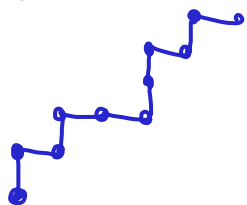
and a sequence $\vec{S}$ is "good" if there is a temperature $\tilde{T}$ ($\beta = \frac{1}{k_B T}$) such that there is a $\tilde{\Gamma}$: $P(\tilde{\Gamma}, \vec{S}) > \frac{1}{2}$ if $T < \tilde{T}$



This model thus allows to explore, in a very simplified way, the relation between sequence and structure.

How many structures (conformations) are there on a square lattice:

- Lower bound: directed walk



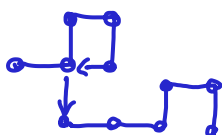
obeys self-avoidance by construction

only upward and rightward

clearly a polymer has more conformations than this, but still they are

$$2^{N-1}$$

- Upper bound: simple random walk with only local self-avoidance



cannot fold back on the previous segment
 $\Rightarrow$ each step has 3 directions

$$\Rightarrow 4 \cdot 3^{N-2}$$

↑
first step

Thus the number N of conformations is

$$2^{N-1} < N < 4 \cdot 3^{N-2}$$

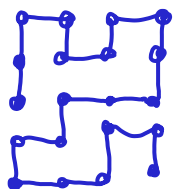
It must grow exponentially:

$$N \propto \mu^{N-1} \quad \text{with } \mu \approx 2.8$$

($\mu \approx 4.6$ in 3D)

The number of conformations is staggering even for short polymers: $N=30 \Rightarrow N \approx 3.3 \cdot 10^{12}$!!!

How many are compact? These are "Hamiltonian walks"



← they fill the space

A rough estimate is

$$N_{HW} \approx \left(\frac{4}{e}\right)^{N-1}$$

First finding: the number of compact conformations grows exponentially too (although being an exponentially decreasing fraction, $\frac{N_{HW}}{N}$, of the total number of conformations).

The problem of finding the good sequences becomes increasingly difficult.

How many possible sequences are there?

$$N_s = 2^N \quad 2 \text{ because } 2 \text{ possibilities}$$

How many of them are good?

We do not have good theoretical arguments $\Rightarrow$ we resort to numerics: find all sequences for which $\exists \tilde{T}$ and T such

$$\text{that } P(\tilde{T}, s) > \frac{1}{2} \quad \text{if } T < \tilde{T}$$

It turns out that $N_{\text{goods}} \sim \gamma^N \quad \gamma < 2$

$\Rightarrow$ grows exponentially, but exponentially vanishing fraction of all possible sequences.

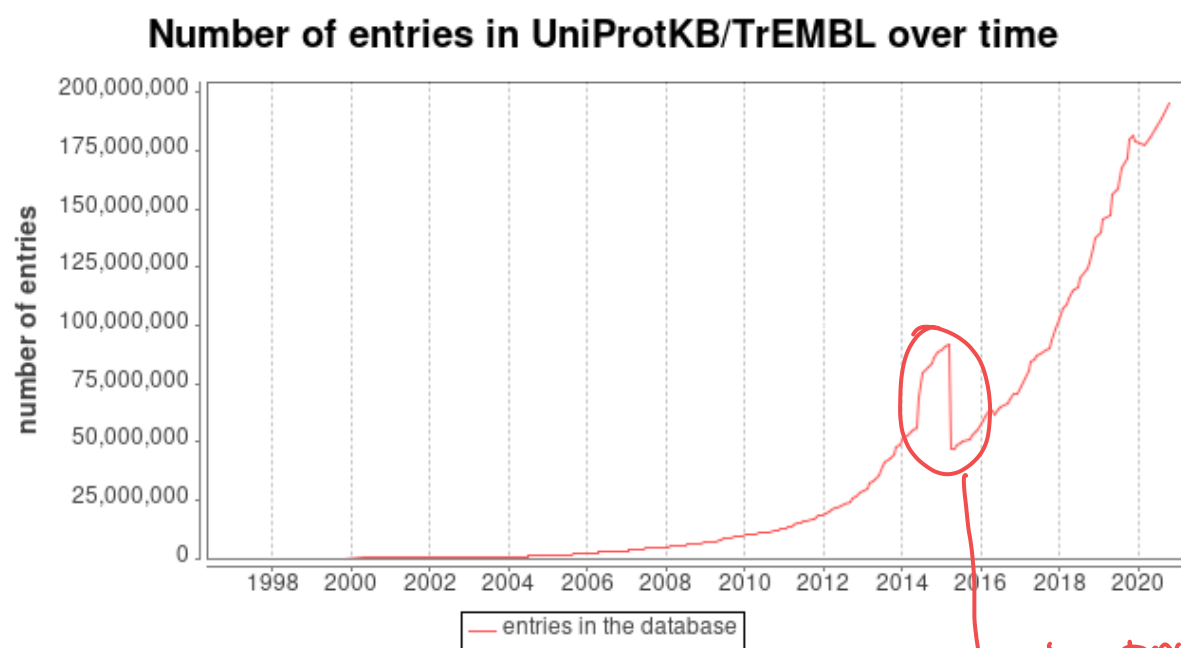
Computational results also revealed something else:

the number of compact structures acting as ground-states (i.e. non-degenerate conformations of lowest energy) grows more slowly than N_{goods} !!!

This means that the same compact structure can be the ground state for many different sequences.

What is remarkable is that this is precisely what has been found for real proteins

Number of reported sequences over time

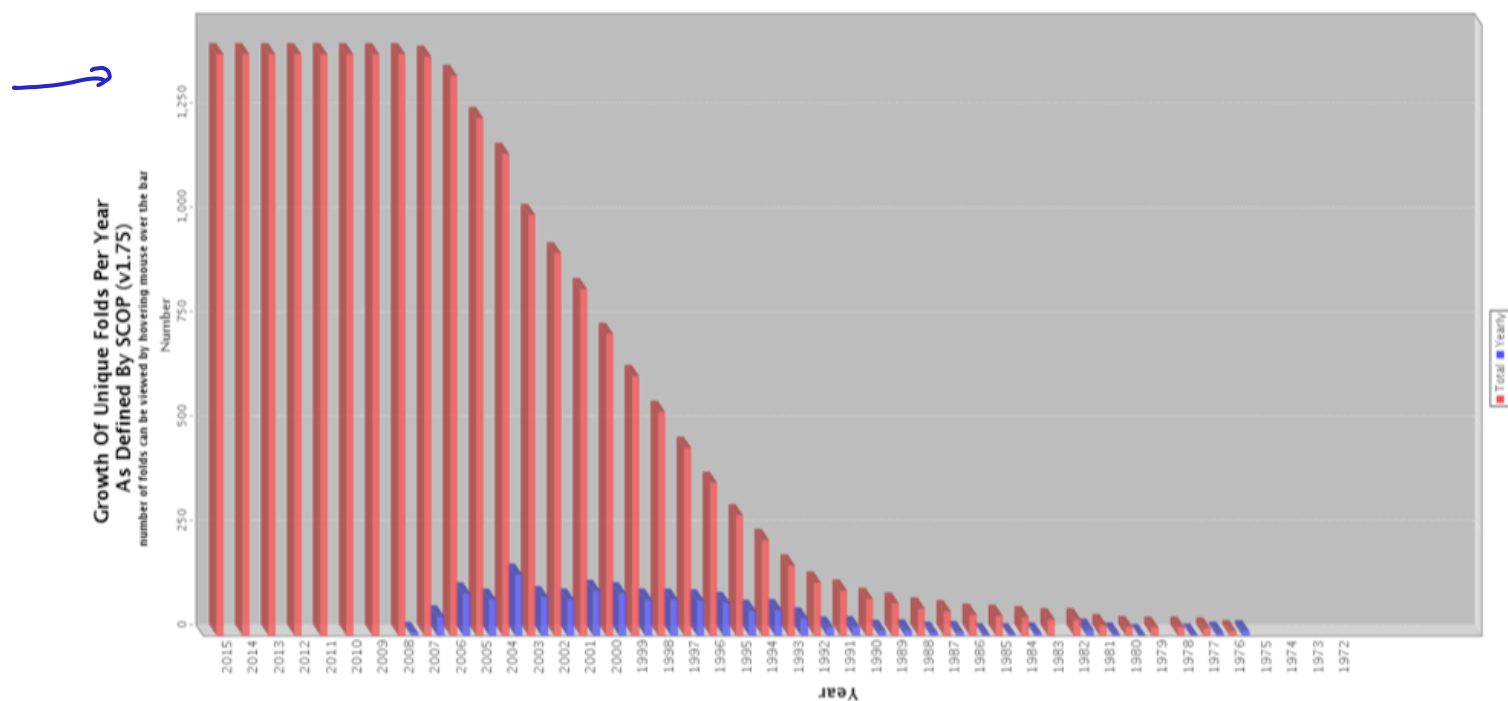


← $200 \cdot 10^6$!!!

→ pruning the database of redundancies

Number of new folds over time :

10^3



It is pretty clear that, as the number of sequences increases exponentially, there is no corresponding increase in the number of new structures. More and more sequences find their ground state over a restricted number of shapes (called "folds")

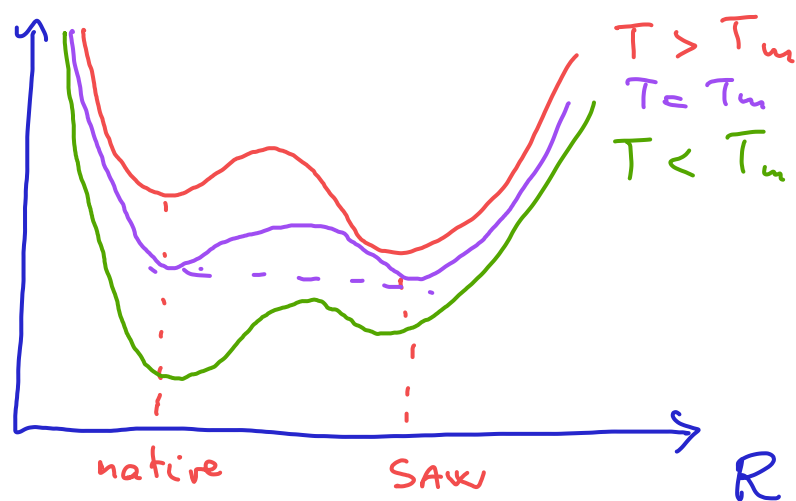
Of course, simple models like the HP model do not describe real proteins, which are way more complex.

For one, they fail at reproducing a crucial feature of most proteins, namely their "two-state" folding property.

Extensive analysis of the "denaturation" of proteins shows

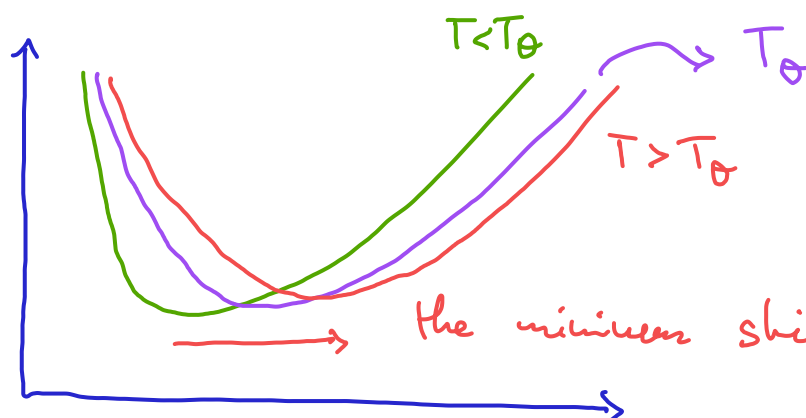
that, although there is a denaturation temperature T_m (m for melting) where $P_{\text{native}} = 1/2$, in proximity of it the protein is indeed found in 2-states: either native, compact and globular, or a random coil (SAW), with very little in between.

From the point of view of the free-energy this corresponds to



At T_m the free-energy (and thus, roughly, the probability) of being native or SAW are the same, shifting correspondingly in the other cases.

In the Flory-de Gennes model, instead,



- For proteins, we talk about :
- two-state transition,
 - all-or-none
(the protein is fully native or not native at all, nothing in between)
 - cooperative phase transition
 - coexistence

But the θ -point transition is a continuous transition, with gradual changes of conformation, rather than two-states.

At present, this behavior is not reproduced by any model with two-body interactions.

Only atomistic simulations including water do.

Otherwise, needs to introduce ad hoc many-body interactions.

