Reading for Lectures 28—29: PKT Chapter 9

The central issues of protein folding surround determination of the full compact global structure.
Biologically relevant proteins fold reproducibly into unique 3D shapes.
The driving force for the folding is energy lowering.
The unique shape is the ground (lowest-energy) state (called the "native state").

The basic problem here is the relation between a 1D amino-acid sequence and a 3D shape/structure. This problem has two versions:
$1D$ sequence $\rightarrow$ $3D$ shape (the "folding problem"), i.e., given the sequence, how to predict the shape.
$3D$ shape $\rightarrow$ $1D$ sequence (the "design problem"), i.e., given a shape, how to find a sequence that will fold to it.

You might think that these are simple problems with a nice one-to-one relation between the two sides. This is not so:

- Not all sequences have a well-defined ground state. (Nature/evolution has selected sequences that do.)
- Not all shapes have sequences that will fold to them (and some "highly designable" shapes have many sequences!) This is a natural limit on the shapes that nature can use for catalytic properties.
- Even if there is a unique ground state, the system has to be able to find that ground state in a finite amount of time. (Here, again, nature has selected sequences for which this is possible.)

Comment:
Important experimental fact: Many (but not all) proteins have the property that they fold spontaneously: If you denature them (e.g., by heating) and then cool back down, they will return to their original shape reasonably quickly (0.1 s—10^3 s) and without special assistance.
But, some proteins require enzymatic assistance (in the form of other protein complexes) to get folded correctly. "Chaperonins" and other special enzymes, including membrane-bound proteins.
Present understanding is incomplete: it is NOT now possible to predict with certainty starting from the 1D sequence:

- What the native state is (or if there is one)?
- And, if there is, how long (and by what route through configuration space) it gets there.

This is a key problem and a lot of work is being done on it. Some progress is being made.

Conceptually the problem is not difficult:
"All" you need to do is to enumerate all configurations (microstates), to have a way of assigning an energy E_n to each one, and to sort through the list to find the lowest energy.

One simple way you can imagine doing this is to label the configurations by the Ramaachandran-plot angles $\{\varphi_n,\psi_n\}_{n=1}^{N}$ and to write down an energy function of those angles $E\left(\{\varphi_n,\psi_n\}_{n=1}^{N}\right)$ including both short and long-range interactions.

The problem is that there are too many configurations: Suppose each of these 2N variables could take only two values. There would then be $W_N = 2^{2N}$ states. For each one, you need to evaluate the energy E_n. Even for a short (e.g., 50-amino-acid protein), that makes $2^{100} \approx 10^{30}$ states. Suppose you examined one state every 10^{-12} s . That would still mean 10^{18} s. At 1 yr = 3×10^7 s, that's 3×10^{10} yrs, compared to the age of the universe at 4×10^9 yrs. Not feasible!

In a slightly different form, this is called "Levinthal's paradox":
How does the protein find its own ground state? (which it does in 0.1--1000 s!)
Where does the 10^{-12} s come from?:

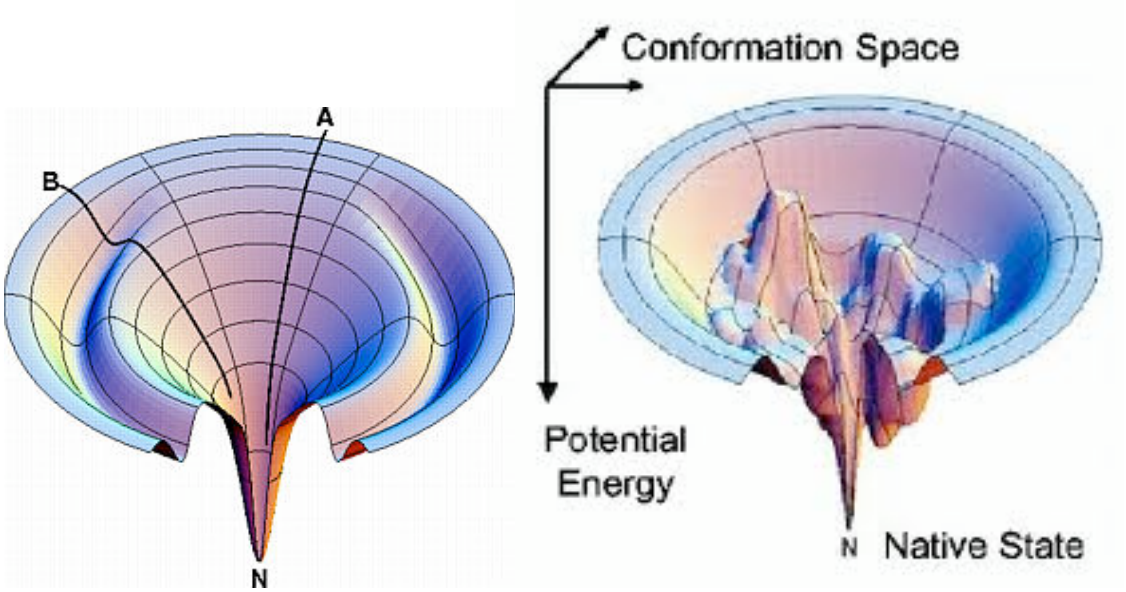
How long does it take to move a distance of 1 nm? $\tau = \frac{d}{v}$, where d= 1 nm and v is a characteristic velocity. Use velocity from thermal energy:

$$k_B T = \frac{1}{2} m v^2 \Rightarrow v = \sqrt{\frac{2k_B T}{m}} = \sqrt{\frac{2\left(1.4 \times 10^{-23}\right)300}{1.67 \times 10^{-27}}} = 2 \times 10^3 \; m/s,$$

where I have used smallest mass (H) to get largest speed and, therefore, smallest time.

Thus, $\tau = \frac{d}{v} = \frac{10^{-9}}{2 \times 10^3} = 5 \times 10^{-13} \; s$.

The resolution of this paradox lies in the concepts of energy landscapes and funnels:
Imagine plotting the energy E_n as a function of a set of coordinates describing the various configurations. Of course, there are a very large number of coordinates; but, to keep things simple, I am going to picture it as if there were only 2.
How does a protein find its ground state? (assuming that it starts from a random-coil configuration)
If you are lucky, then you don't have to explore the whole of configuration space. You just follow the gradient down to the lowest point, which is the "native" state.
It is believed that this is what happens in some favorable cases.
But, there are alternatives. Suppose that the landscape is "rougher", so that there are places where the protein can get stuck in a local minimum. This also happens in some cases, and that's a place where the chaperonins come in to help by "rescuing" folding from local minima.



Note: Importance of $_{kBT}$ scale on plots above.
But, this is by no means to worst case:
Suppose that the energy landscape has many deep minima, so the folding can get started in an entirely wrong way?
Connection to prion diseases.
Or, suppose that it is very flat, with a very small basin of attraction of the native-state funnel?
Present understanding is that any and all of these things can happen.
Nature/evolution has selected proteins which fold reliably and rapidly.
The other ones are out there; nature just does not use them (for the obvious reasons).

The present understanding of the folding process is that—for the proteins that occur naturally—there are two/three steps:

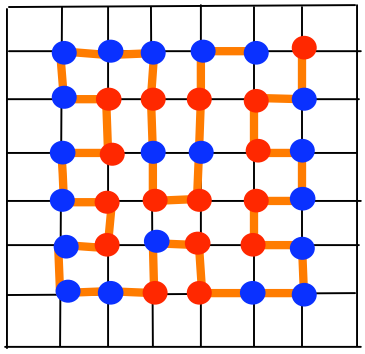
- Quick hydrophobic collapse into a "molten blob."
- Slower thermal refolding down into a single funnel.
- Possibly, rescue from common traps by chaperonins.

There is a lot of work—essentially all numerical—in dealing with the energy landscape, the energy spectrum, and the folding process.

Models include:

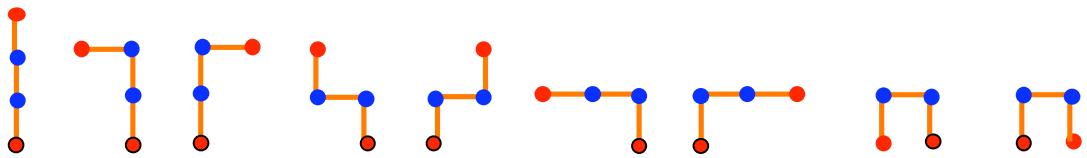
- Ramachandran-angle models and other off-lattice models, which are trying to deal with the "real" world. Computationally very intensive.
- Lattice models, which try to simplify as much as possible. Not realistic, but computationally tractable-enough to look at conceptual issues.

Lattice models: HP models



Here's an example of a compact (e.g., molten-blob) model protein.
It is common to divide all the 20 amino acids into only two groups polar (P=blue) and hydrophobic (H=red).
The 1D sequence of this 36-mer starting from the upper left corner is:
HPHHPPHHPPPHHPHPPPHHPPHHPPPPPHPHHPHPP
We then need to assign an energy to this.
Two kinds of energy:
Interaction energies (nearest-neighbor only): ε_{HH}, ε_{PP}, ε_{HP}.
Solvation energies: ε_{WH}, ε_{WP}
Water-Water: ε_{WW}
It would be common to take:
$\varepsilon_{WW}=0$ (just sets an irrelevant energy scale)
$\varepsilon_{WP}<0$ (polar residues like to dissolve, favored for outside of cluster)
$\varepsilon_{WH}>0$ (hydrophobic residues don't like to dissolve, favor inside of cluster)
$\varepsilon_{HH}<0$ (hydrophobic residues attract), etc.
Note: We do NOT count the (fixed) covalent contacts along the backbone.
In this example: $E_n = 16\varepsilon_{WP} + 4\varepsilon_{WH} + 10\varepsilon_{HH} + 9\varepsilon_{HP} + 6\varepsilon_{PP}$
Your strategy is to enumerate all the configurations along with their energies.
Find the lowest-energy (native state) and other low-lying states.
Calculate the equilibrium probabilities according to $P_n = \frac{1}{Z} e^{-\beta E_n}$.

Here's a trivial example: Consider the tetramer HPPH on a square lattice of N sites:
There are 4Nx9 different configurations:

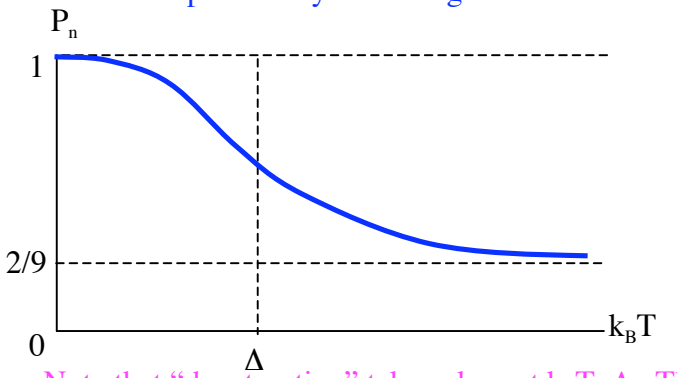


The first 7 of these 9 are "extended" and have the same energy: $\varepsilon_{ex} = 4\varepsilon_{WP} + 6\varepsilon_{WH} \equiv \varepsilon + \Delta$
The last two are "compact" and have energy: $\varepsilon_c = 4\varepsilon_{WP} + 4\varepsilon_{WH} + \varepsilon_{HH} \equiv \varepsilon$
The energy difference $\Delta \equiv 2\varepsilon_{WH} - \varepsilon_{HH}$ is positive because (a) $\varepsilon_{HH} < 0$ and (b) $\varepsilon_{WH} > 0$.

In terms of our terminology, the HPPH sequence has a unique "native" (ground) state (up to rotations, translations, reflections).

In a thermal ensemble: $Z = 7e^{-\beta(\varepsilon+\Delta)} + 2e^{-\beta\varepsilon}$

- the probability of finding the native state is $P_n = \frac{1}{Z}2e^{-\beta\varepsilon} = \frac{2}{7e^{-\beta\Delta}+2} = \frac{2e^{\beta\Delta}}{7+2e^{\beta\Delta}}$
- the probability of finding an extended state is $P_{ex} = \frac{1}{Z}7e^{-\beta(\varepsilon+\Delta)} = \frac{7e^{-\beta\Delta}}{7e^{-\beta\Delta}+2} = \frac{7}{7+2e^{\beta\Delta}}$

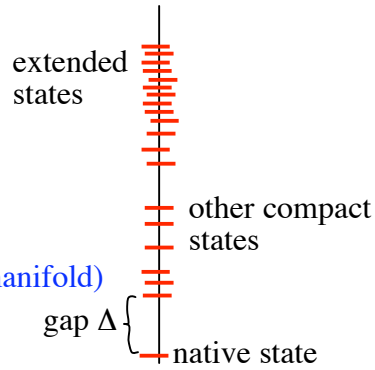


Note that "denaturation" takes place at $k_BT \sim \Delta$. The transition is sharper for large N because there are an increasingly large number of the exentended/random-coil states.
(I will give you a problem of this type on the next homework)

What are the messages of these studies of model systems?:

A. Folding Problem Results:

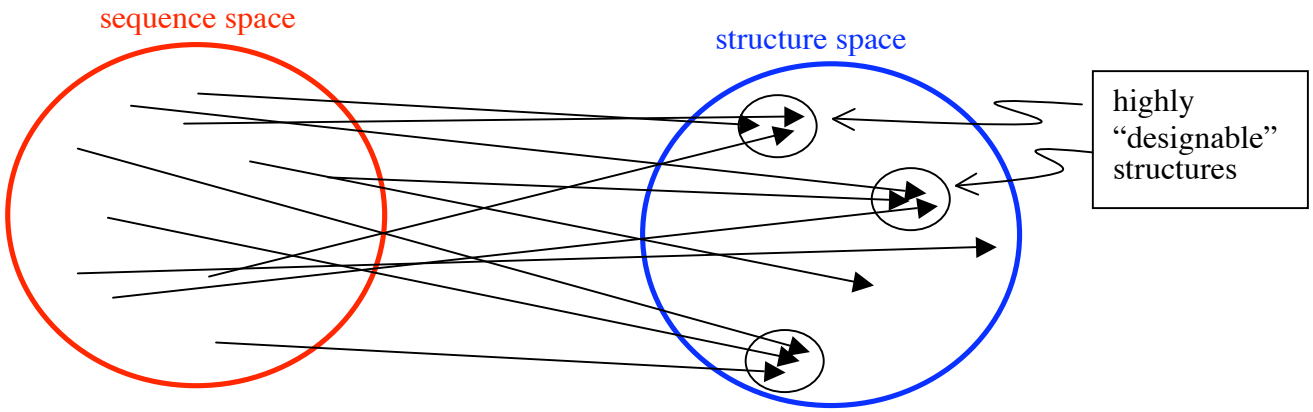
Low-lying states are compact, assuming significant number of H residues.
Most sequences do not have a unique native state (or "manifold").
Most sequences have a "rough"energy landscape.
Most sequences do not have a well defined "folding funnel".
Some "good" sequences do have an unique native state.
These are often characterized by a large "gap" between the ground state (or manifold) and the first excited state. "themodynamic stability"
These "good" states tend to have a robust folding funnel.



B. Design Problem Results:

Most compact structures have no (or very few) sequences which fold to them.
A few structures are highly "designable" in the sense that many sequences fold to them.

The energy function provides a mapping from sequence space to structure space:



If you understand this mapping, you can hope to design "new" folds/shapes, not used by nature/evolution.

Overall message: Nature/evolution has chosen "designable" shapes, selected (or course) to be useful either structurally or enzymatically.