The protein folding problem consists of two parts:

1) Creating a stable, well-defined structure that is significantly more stable than all other possible structures.

2) Finding a way to get to that structure
Why start a lecture on protein folding with a picture of DNA. The reason is that protein structures are very different from the structure of DNA and by understanding what is so different between the two will help us better understand the essence of the protein folding problem.
The structure of myoglobin was a complete shock for people who had hoped that the structure, once revealed, would immediately explain how this molecule worked.

The structure was startling in two ways

1) The structure was not regular at all. Yes there were helices, but they were arranged in what appeared to be random orientations.

2) Even though the structure was completely irregular, the protein had a defined structure. This was even weirder. How can a structure be so “random” but at the same time be so well defined.

Obviously understanding how proteins worked would be a lot harder than understanding
 How DNA worked.
Comparison to other self-assembling systems?

In this respect proteins are different from other polymers like DNA. These polymers also adopt a defined structure by finding the global energy minimum but this global energy minimum is achieved by finding local minima for parts of the structure. In other words we perform multiple independent searches. Dividing up the search for the global minimum global energy minimum by finding the local conformational minimum for each part
The search for the conformational energy minimum is global
not
a combination of multiple independent local searches

Traveling salesman problem a classic example
of a global search problem

The conformational search problem for proteins are different from other polymers like DNA. These polymers also adopt a defined structure by finding the global energy minimum but this global energy minimum is achieved by finding local minima for parts of the structure. In other words we perform multiple independent searches. Dividing up the search for the global minimum global energy minimum by finding the local
The conformational energy landscape describes the relative energy of all possible conformational states of a molecule.

Two-dimensional energy landscape

The conformational energy landscape of a protein describes the relative energy of all possible conformational states. Above is a two dimensional energy landscape. The z (vertical) axis give the relative energy. The x and y axis represent two parameters that describe the structure. This could be torsion angles for example. So if we had a molecule that had only two torsion angles then the energy landscape may look like the one in the picture. The location of the sphere represents the current conformational state of the molecule. The energy landscape will have little divots, local minima and the molecule could become trapped in these minima. However, there is thermal energy that constantly jostles the molecule around, so the molecule changes conformations i.e. the sphere moves around and can “explore” the energy landscape.

To describe the conformation of a protein we obviously need a lot more parameters than two. So the protein energy landscape has many different dimensions. The backbone alone contributes two dimensions (phi, psi) for every peptide unit. Add to that bond stretching, bending of angles etc. and the energy landscape of a 150 AA protein can easily have hundreds or thousands of dimensions. We obviously cannot draw such an energy landscape, but the concept of energy landscapes is very useful when thinking about protein dynamics and so researchers will often draw a one- or two-dimensional energy landscape to make a point.
Frustrated systems
The thermodynamic hypothesis

The native structure of a protein is determined solely by the sequence of amino acids in its poly-peptide chain and represents the state of the lowest conformational energy under native conditions.

Christian Anfinsen

Nobel Prize in Chemistry 1972

There are two important points to this statement:

1) The sequence determines the structure. Yes, chaperons help, but they do not alter the structure. Instead of shaping the structure, these chaperons appear to help the protein’s get to that structure.

2) The native structure is the state of the lowest conformational energy. Because that energy minimum is unique, the structure is unique. Most polymers, including random polypeptide chains do not follow the thermodynamic hypothesis. Instead the structure of those polymers is usually determined by kinetic factors. The polymer gets stuck in some local conformational state.
Solving the energetic problem

Make more stabilizing interactions than you break

- **Hydrophobic effect**
- **Conformational entropy**
- Hydrogen bonds
- Electrostatic interactions
- + Van der Waals interactions

very small Number

The main driving force towards a folded structure is the hydrophobic effect. The main driving force opposing folding is the loss of conformational entropy.

Let's estimate the size of some of these individual interactions.
Another physical restriction on domain structure is protein size. One of the dominant energetic trade offs between the folded and unfolded state is between conformational entropy in the unfolded form and hydrophobic stabilization in the folded form. Remember both of these effects are entropic in nature.

The per-residue entropic cost of folding is independent of the proteins size. The total entropic cost of folding a protein is simply the sum of the entropic cost of restricting the conformational degree of freedom for each amino acid.

By contrast the amount of entropy we gain through the hydrophobic effect depends on the proteins size. The larger a protein gets, the larger the fraction of its sidechains that it can bury in a hydrophobic core. This is simply due to the fact that the volume of the protein increases with the cube of the radius while the surface area increases with the square of the radius.

Does the combination of a size-independent entropic cost with a size-dependent entropic gain for the folding reaction impose any restrictions on domain size? Yes, because if we make our domain smaller and smaller, we have fewer and fewer hydrophobic interactions to pay for the cost of constraining the conformational freedom of our sidechains.

So what fraction of a protein’s sidechains has to be in the hydrophobic core to give the folded state a net stability?
Entropic cost per amino acid $0.6 \text{kcal/mol} \times \ln 10 = 1.3 \text{ kcal/mol}$

Average number of methyl groups per hydrophobic amino acid 5

Hydrophobic effect per buried amino acid $= 5 \times 0.8 \text{ kcal/mol} = 4 \text{ kcal/mol}$

--- $\sim 1/3$ of amino acids need to be buried

$$V = \frac{4}{3} \pi r^3 \quad \frac{4/3 \pi x^3}{4/3 \pi y^3} = \frac{\text{hydrophobic}}{\text{total}}$$

$x=2 \ ; \ y=4 \quad n=64 \quad \text{ratio } = 8$

$x=3 \ ; \ y=5 \quad n=125 \quad \text{ratio } = 4.6$

$x=4 \ ; \ y=6 \quad n=216 \quad \text{ratio } = 3.4$

Lets do a very simple minded calculation. Lets say that we restrict a sidechain’s conformational freedom by a factor of 10 when we go from the unfolded form to the folded form. So our entropic price per residue is $\sim 1.3 \text{ kcal/mol}$. The average hydrophobic sidechain has 5 carbons and the average energy of burying a methyl group is 0.8 kcal/mol to burying a side chain should give us $\sim 4 \text{ kcal/mol}$. So in order to get net stability from folding we need to bury approximately 1 out of every three amino acid sidechains in a protein’s hydrophobic core. So the volume of the hydrophobic core needs to be greater than 1/3 the volume of the total protein. It turns out that we get a ratio of $\sim 3$ for domains on the size of 125 - 200 amino acids, which is close to the lower size limit for domain size. There clearly are some domains that are smaller than this size limit, but there are very few domains consisting of less than $\sim 80$ AA.

Another way of thinking about this is if we have a cylindrical protein shape and we take the middle figure of the previous slide as a cross section we need at the very least a double layer of hydrophobic sidechains in the interior of the protein. To get close to this $1/3$ ratio.
Here is an example of the helical bundle fold in which we have this double layer of hydrophobic residues in the interior of the protein.
Here is the example of the ribonuclease inhibitor which again shows this double layer of hydrophobic residues.
There are some protein domains that are much smaller than our size limit of ~80 amino acids, but these domains are not stabilized by hydrophobic interactions alone. Instead these domains require some other source of stabilizing energy such as the binding of a metal ion or cysteine bridges. To stabilize the structure

Protein engineers have actually been able to generate zinc finger proteins like the one shown on the left above, that were stabilized by hydrophobic interactions alone. However in such a “domain” each interaction is absolutely essential to keep the structure intact. Such a domain would have very little chance of being retained by an organism over the course of evolution. One deleterious mutation would lead to a complete loss of the domain’s structure and function and it is very unlikely the gene would acquire a compensatory mutation before the gene is lost, or receives a second deleterious mutation. In other words, protein domains that are very small, probably do not show the degree of robustness that is needed to retain protein structure in the face of random mutations.
Using the hydrophobic effect to evade immune Detection: Neisseria gonorrhoeae pilin

Parge et al. 1995 Nature
Using the hydrophobic effect to evade immune detection: Neisseria gonorrhoeae pilin
Levinthal’s paradox

Proteins have an incredible number of possible conformational states, yet they are able to fold very quickly.
Levinthal’s paradox

150 AA domain
10 different conformations per side chain
------> $10^{150}$ possible conformations.

Atomic vibrations occur on fsec. ($10^{-15}$ sec) time scale

Time to search all possible conformations >> age of the universe

Number of particles in the universe $\sim 10^{85}$
**Levinthal’s paradox**

Random protein sequences typically do not fold, neither do most other polymers.

There must be a code / mechanism
That allows proteins to fold within a reasonable Period of time.

Two models:

Frame work

Global collapse
Framework model

Secondary structure elements form first
Packing of secondary structure elements leads to molten globule
Repacking of the core in the molten globule leads to native structure
**Global collapse model**

- Protein collapses due to hydrophobic effect
- Hydrophobic environment drives formation of secondary structure to form molten globule
- Repacking of molten globule leads to native structure
Phi value analysis

Alan Fersht

\[
\Phi = \frac{\Delta \Delta G_{o,U}}{\Delta \Delta G_{o,F}}
\]

Phi value =1  interaction already exists in transition state
Phi value =0  interaction is not part of transition state
Current state of debate

A detailed understanding of protein folding remains illusive because we still lack experimental information on many of the states along the folding trajectory.

The transition state of a two-state folder tends to be very compact.

Proteins with similar folds tend to fold following a similar mechanism.