How Does a Protein Fold?

In this section we consider the problem of how a polypeptide folds into a compact, active, globular protein, with its 3d structure determined by its sequence. We first review the famous *Levinthal paradox* which presents a very pessimistic viewpoint suggesting that proteins will never find their native structure because of the overwhelming entropy of nonfolded states.

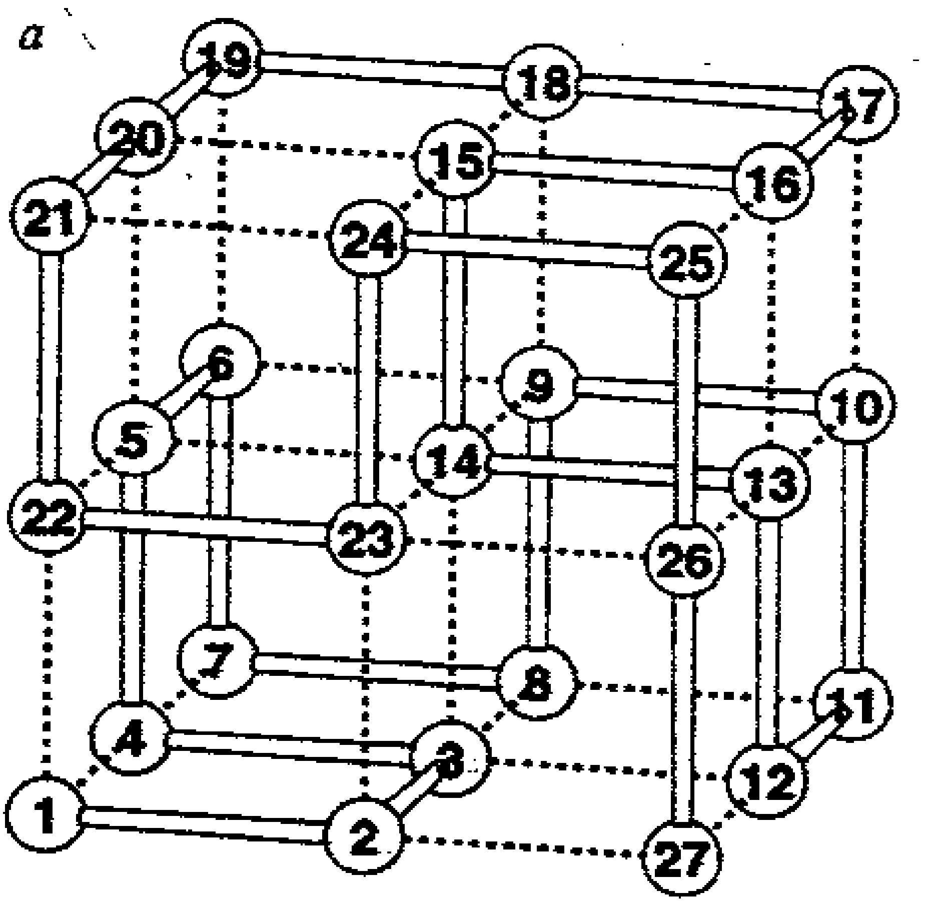
We then describe how this paradox can be resolved by recognizing the entropy-reducing effect of compaction of the protein into a disordered random coil, and then finally we describe the importance of sequence in defining a well-defined native state.

A lot of the ideas discussed below are pulled from a very clear paper by Andrej Sali, Eugene Shaknovich and Martin Karplus (Nature 369, 248 (1994)).

A Simple `Lattice' Model of a Protein

Consider an N-amino acid polypeptide. For simplicity let's think of it as a `polymer' on a 3d cubic lattice. Each amino acid occupies one site on the lattice, and each peptide bond sits on a bond of the lattice. Although a lot of chemical detail is ignored in this model, it provides a nice way to count conformational states.

The figure below shows one compact (3x3x3) configuration of a 27-amino-acid chain for this idealized model.



[Figure from Sali et al, Nature 369, 248 (1994)]

Number of States

An important question is how many conformational states there are for our N-vertex polymer.

Denatured (Random Coil) Protein

Ignoring self-avoidance and assuming the protein to be fully denatured (unfolded) there are $W \approx 6^N$ states, since each peptide bond can be put along one of 6 possible directions on the lattice.

[purists might want to use N−1 instead of N as the exponent - this will not be critical for our purposes]

However, this estimate includes sharp `hairpin' bends where the peptide backbone reverses direction. Throwing these configurations out gives $W \approx 5^N$ `denatured' states.

Finally, you might want to enforce self-avoidance, and allow each node of the lattice to be occupied by at most one amino acid. Now you have to do a numerical calculation to find that the number of states is further reduced to $W \approx N^{4.7}$, really just a bit less than our no-backtrack estimate.

[see de Gennes book, p. 39 for the details of the number of states in the self-avoiding case]

Now think of a `typical' denatured protein, of $N \approx 100$ amino acids. The number of denatured states is $W \approx 4.7^{100} \approx 10^6 7$.

You can already see a problem - there are a lot of unfolded states.

Levinthal Paradox - You'll Never Find the Native State by a Random Search

The dumbest possible estimate of the time it will take a protein to fold is to suppose that by random thermal motion, one state is `visited' every time interval τ_0. The time needed to fold is therefore

$$t_{\mathrm{fold}} = W \tau_0$$

Given the estimate of $W \approx 4.7^N$ discussed above, we just need τ_0.

Just for argument's sake, let's suppose that τ_0 is the time needed for one segment of the polypeptide (of length $b \approx 3$ Å) to diffuse a distance of its size, or

$$\tau_0 = \frac{\eta b^3}{k_B T}$$

where η is the viscosity of water. Plugging in $\eta = 0.001$ kg/(m sec), $k_B T = 4.1 \times 10^{-21}$ J and $b = 3 \times 10^{-10}$ m gives $\tau_0 \approx 10^{-11}$ sec or about 10 picoseconds.

[we could guess that the `real' value of τ_0 should be quite a lot longer since conformational diffusion will not renew the molecule conformation very efficiently, and will instead stick near the same overall molecule shape for quite a bit longer than 10 psec - again this is not central to what will follow]

This gives $t_{\mathrm{fold}} = 4.7^N \times 10^{-11}$ sec. If we plug in $N = 100$, we obtain $t_{\mathrm{fold}} \approx 10^{57}$ sec. This is a long time (note that one year is 3.16×10^7 sec).

The conclusion of this kind of line of reasoning - that the time needed to visit all the states of a protein and therefore to `find' its native folded state, is 10^{30} times longer than the expected lifetime of the universe (much longer than the few seconds observed experimentally) - is often called the Levinthal paradox.

The famous original reference is C. Levinthal, in Mossbauer Spectroscopy in Biological Systems (ed. P. Debrunner, J.C.M. Tsibris and E. Munck) 22-24 (University of Illinois, Urbana IL 1969).

Resolving the Levinthal Paradox I - The Molten Globule Has Fewer States than the Fully Denatured Protein

What is wrong with the Levinthal estimate? The first thing is that the estimate of $W \approx 4.7^N$ is appropriate for a random coil polymer, and not for a protein in the early stages of its folding, which will experience attractive self-attractions. A protein can certainly not be expected to fold under conditions that keep it in an entirely denatured state.

In fact, when a denatured protein is dumped into normal solution conditions, the hydrophobic amino acids quickly glob together, making the polypeptide have a structure similar to the collapsed globule state of a polymer that we discussed previously. The main difference is that the exterior of the protein has a bunch of hydrophilic amino acids on it.

So, instead of $W \approx 4.7^N$, we should use the much smaller number of `collapsed globule' states, the number of ways to draw our protein on our lattice so that it is collapsed. Computer studies have shown that in this case we have instead,

$$W \approx 1.85^N$$

see Pande et al, J. Phys. A 27, 6231 (1994). This is a much smaller number of states than for the self-avoiding denatured case.

For N = 100 our revised estimate gives $W \approx 10^{27}$, unfortunately still requiring 10^{10} years to fold. Although this is now comparable to the lifetime of the universe, the physics is not right yet.

Thinking along the lines discussed above - considering only the entropy (number of states) - is fruitless because W will inevitably be exponentially dependent on N (i.e. the entropy will end up being extensive in N). What is missing is any consideration of the detailed energetics of the problem that define the native state as thermodynamically favorable, i.e. the amino acid sequence.

Problem: Very roughly explain why $W \approx 2^N$ for the compacted random globule state.

Hint: Think about the number of ways to winding a string into a ball.

In case you are interested, on the 2d square lattice, the number of ways to draw a compacted polymer is about $W \approx 1.47^N$, see Jacobsen et al, Nucl. Phys. B 532, 635 (1998).

Resolving the Levinthal Paradox II - Amino-Acid Sequence

From now on we consider only the collapsed, random globule state, with $W \approx 1.85^N$.

Let's put sequence on our protein. Again opting for the simplest possible model, we choose only two types of amino acids, H and P. We suppose that adjacent H-H and P-P contacts have an energy of $-\varepsilon$, while adjacent H–P contacts have an energy of $+\varepsilon$ (where $\varepsilon > 0$).

We can keep in the back of our minds that ε for real proteins is on the order of k_B T at room temperature, i.e. $\varepsilon \approx 10^{-20}$ J.

This simplistic two-amino-acid model is actually used in real research on fundamental statistical mechanics of proteins.

For further simplicity let's consider the case where the numbers of H's and P's along the polypeptide are equal.

Now let's estimate the `density of states', or the number of configurations per increment in energy, or dW/dE. For the model as described above, the distribution of energies should be symmetric about E = 0, and therefore has mean energy zero.

Since there are on the order of N contacts between amino acids in each random globule state, it is reasonable to suppose that the distribution of states is a Gaussian distribution with energy-width $\approx N^{1/2}\,\varepsilon$ (think of the summing of interactions along the chain as a random walk in *energy*). After properly normalizing the distribution we have

$$\frac{dW}{dE} = \frac{z^N}{(2\pi N\,\varepsilon^2)^{1/2}} \exp[-E^2/(2N\varepsilon^2)\,]$$

where z = 1.85. This distribution is symmetric about E = 0 with zero mean energy, and has the right width and normalization ($\int dE\,(dW/dE) = W = z^N$). This assumed distribution is reasonable, but can only really be backed up by computer calculation.

The main properties of the distribution are:

1. The most probable states have energies with E = 0;

2. Typical states have energies between $-N^{1/2}\varepsilon$ and $+N^{1/2}\varepsilon$;

3. The lowest energy states have energies ≈ $-N\,\varepsilon$ (i.e. all contacts generating energy $-\varepsilon$.

This last property can be shown to be consistent with the assumed distribution by determining at what energy $dW/dE \approx 1/\varepsilon$. At this point the energy spectrum is sufficiently sparse that there is just one state per energy increment.

Problem: Find the energy at which $dW/dE = 1/\varepsilon$.

We'll assume that the lowest-energy state is at energy $-E_0 = -c\,N\,\varepsilon$ where c is a numerical constant (i.e. some number not too far from 1 in magnitude). The precise value of c will depend on the exact sequence.

Finally note that the *highest* energy state must have energy $+E_0 = +c\,N\,\varepsilon$.

Therefore we can write the energy density of states more precisely as:

$$\frac{dW}{dE} = \begin{cases} \dfrac{z^N}{(2\pi N\,\varepsilon^2)^{1/2}} \exp[\,-E^2/(2N\varepsilon^2)\,] & |E| < E_0 \\ 0 & |E| > E_0 \end{cases}$$

since there are no states with energies outside the range $-E_0$ to $+E_0$.

What is nice about this model is that is very amenable to statistical-mechanical analysis. All we need to do is tack on a Boltzmann factor to find the probability distribution for states of energy E at any temperature:

$$P(E) \propto \frac{dW}{dE} \exp[-E/(k_B\,T)] = \frac{1}{(2\pi N\,\varepsilon^2)^{1/2}} \exp[\,-E/(k_B\,T) - E^2/(2\,N\,\varepsilon^2)\,]$$

for $|E| < E_0$ (and zero outside of this range).

Now, it is no surprise that for sufficiently low temperature T the *one* lowest-energy state at $E = -E_0$ will become more probable than the ≈ z^N states near E = 0. The ratio of the probabilities of the lowest-energy state and the E = 0 state is just

$$P(-E_0)$$

$$\frac{}{P(0)} = \exp[\ E_0/(k_B\ T) - E_0^2\ /\ (2\ N\ \varepsilon^2)\ ]$$

Plugging in $E_0 = c\ N\ \varepsilon$ gives

$$\frac{P(-E_0)}{P(0)} = \exp\{\ cN\ [\ \varepsilon/(k_B\ T) - c/2\ ]\ \}$$

which tells us the temperature at which the lowest-energy state actually becomes more probable than the states at $E = 0$,

$$T^* = \frac{2\ \varepsilon}{c\ k_B}$$

Above this temperature, the entropic peak near $E = 0$ dominates, and the protein explores many `random globule' states.

Below this temperature the lowest-energy state dominates, and the protein `freezes' (or `folds') into one state.

Note that the temperature range near T^* over which this change occurs is $\approx T^*/(cN)$, and therefore becomes increasingly sharp as N is increased.

There is a kind of phase transition from a disordered globule to the native state at a finite temperature related to the energy of interaction of adjacent amino acids, i.e. around room temperature.

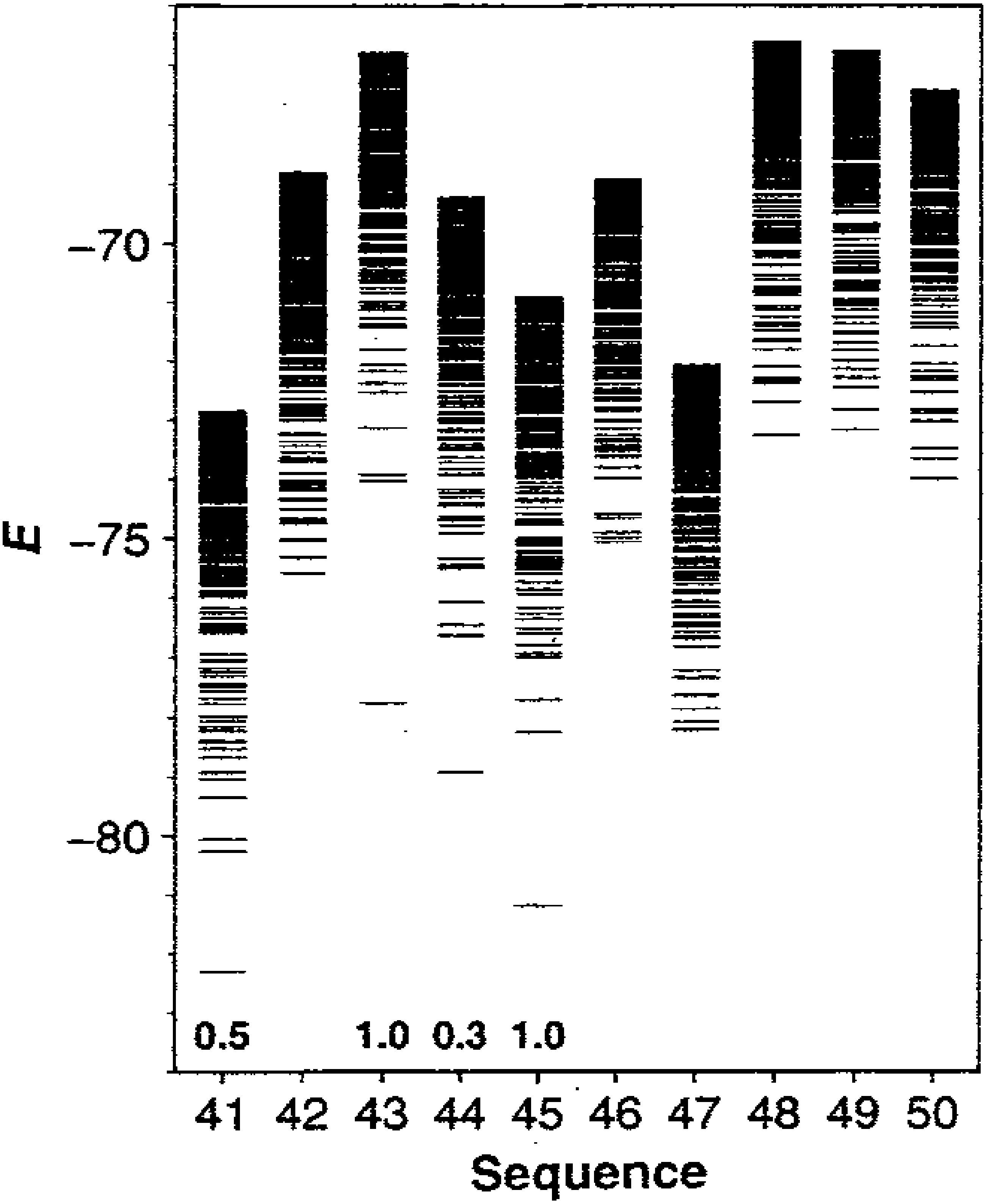
The model discussed above is sometimes called the *random energy model* and was first discussed in a non-protein context in a paper by B. Derrida (Phys. Rev. B 24, 238 1981). Its connections to protein folding has been discussed in work by Grosberg and by Shaknovich.

Resolving the Levinthal Paradox III - Make Sure There is a Native State

We saw in the previous section that there is a kind of freezing that can occur at a finite temperature. Above the freezing temperature, most of the time our model protein is unfolded, with energy near zero. But, below the freezing temperature, the protein will be in a state with energy $\approx -N\ \varepsilon$.

But will it really be folded uniquely? Work of Sali and coworkers showed that for random sequences, roughly like the model described above, there will be a more-or-less continuous distribution of energy levels all the way down to the `ground state', and the freezing will not occur uniquely into one state. Different low-energy states will be reached if you repeatedly unfold and refold the protein by cycling the temperature above and then back below T_m.

Sali and coworkers did find that *some* sequences would repeatedly fold to the same state. You can see a few of these cases in the `energy level' figure in their paper, and these cases have a single low-lying ground state, well separated from the gaussian-like distribution of the rest of the state.

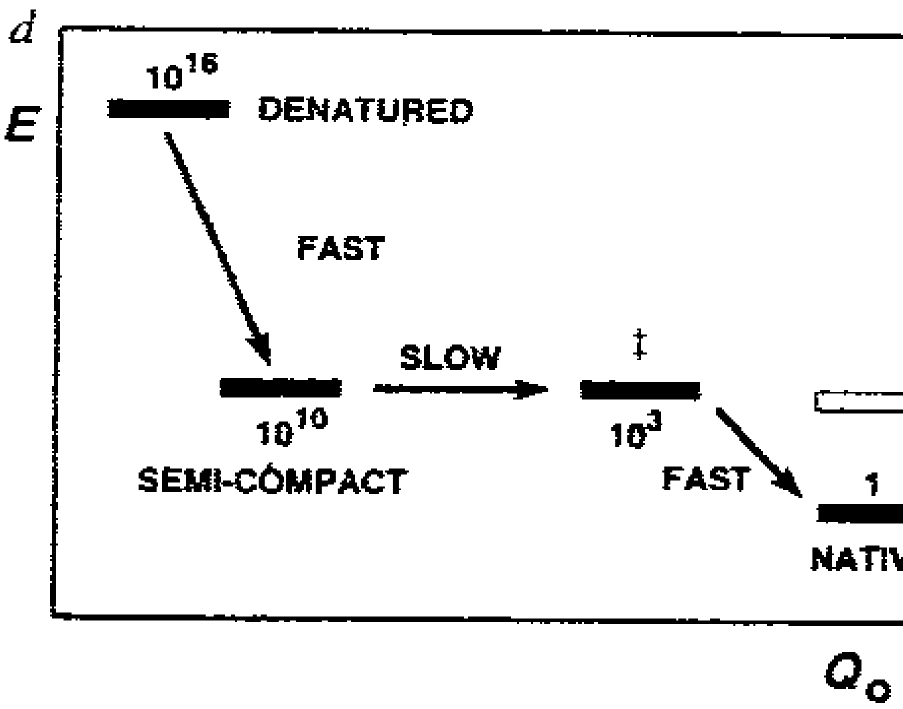


[Figure from Sali et al, Nature 369, 248 (1994); Part of caption for above figure from Sali et al, Nature 369, 248 (1994):
Energy spectra for 10 folding and non-folding random sequences. The energies of the 400 lowest compact self-avoiding conformations are shown. The native state corresponds to the bottom bar. The numbers below the spectra show the folding tendencies of the corresponding sequences. If no number is given the folding tendency is 0. A sequence folds in a given MC simulation if it finds the native conformation within 50×10^6 MC steps. Folding tendency of a given sequence is defined as the fraction of 10 MC runs that started with a random conformation under a given set of conditions. A sequence is a folding sequence if the native conformation is structurally unique and folding tendency is high (≥ 0.4) under conditions where the native structure is thermodynamically stable. A sequence is a non-folding sequence if the folding tendency is 0.0. There are 24 intermediate sequences that we do not consider here.]

Sali et al realized that these special cases (which they found by accident) showed that it was possible to have sequences which `lead' a protein to fold to a unique ground state. They also suggested that proteins with such sequences gave a selective advantage to the organisms that happened to create them, since they had a more reliable native state, and therefore chemical function.

Summary of the Kinetics of Folding of a Protein to its Native State

This figure from Sali et al suggests how the kinetics of protein folding typically works, in cases which do and which do not fold into a unique native state. The vertical axis indicates enthalpy (energy) and the horizontal axis indicates time.



[Figure from Sali et al, Nature 369, 248 (1994)]

File translated from TEX by TTH, version 2.53.
On 30 Mar 2001, 22:53.