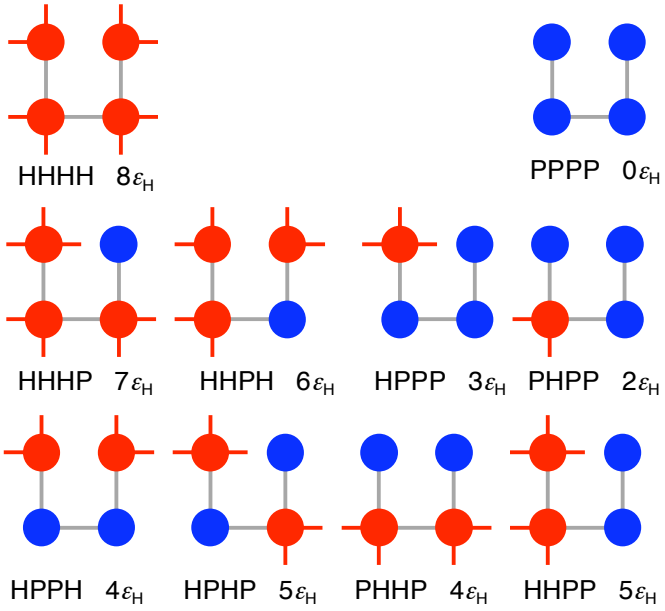
PHYS 4xx Poly 6 - Models for protein folding

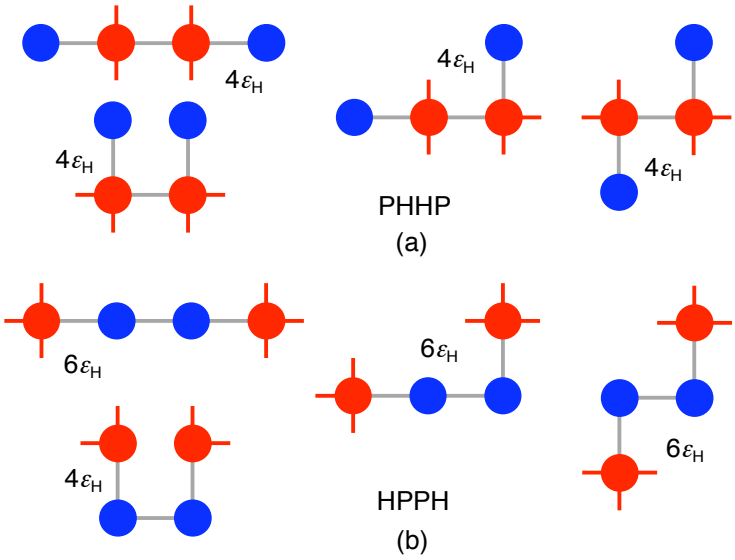
The diameter of a protein increases roughly as the 1/3 power of its contour length, arguing that proteins are densely packed or folded.
We examine a variant of the HP model for protein folding (Hydrophobic/Polar), a one-parameter model (Lau and Dill, 1989) in which every amino acid is classified as either hydrophobic (H) or polar (P): 9 amino acids are H (alanine, cysteine, isoleucine, leucine, methionine, phenylalanine, tryptophan, tyrosine, valine) and 11 AA are P (arginine, asparagine, aspartic acid, glutamic acid, glutamine, glycine, histidine, lysine, proline, serine, threonine).
Here, hydrophobic species have a repulsive energy $+\varepsilon_H$ with any polar species or the environment; all other interactions are zero; nearest neighbor covalent bonds are not part of this accounting; the configurations live on a square lattice in two dimensions (Phillips *et al.*, 2009).

Consider a model polypeptide with just four amino acids. Its lattice conformations are:
*a straight line with four consecutive lattice points
*an L-shape with three lattice points on a line and one point off the axis
*a zig-zag shape with the two end points of the polymer lying on opposite sides of the central bond
*a U-shape (with sharp corners).
Onto these structures are placed $2^4 = 16$ sequences of various combinations of H and P amino acids. The U-shaped configurations are shown below: polar amino acids are colored **BLUE**; hydrophobic ones are **RED**. The short red lines emanating from the hydrophobic sites represent repulsive interactions, each with an energy cost ε_H; the total number of these lines is the repulsive energy of the configuration in units of ε_H.



In the U shape, the H sites have less exposure to the aqueous environment so that the energy of the chain drops as a result. The hydrophobic effect favors folded states.

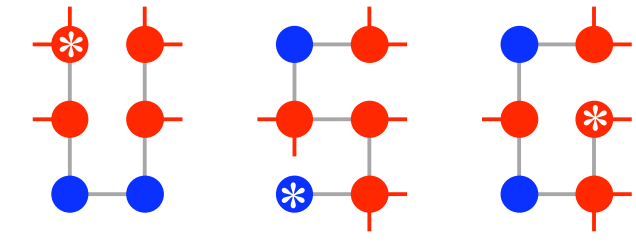
We search for the spatial structure that minimizes the energy for a given HP sequence. For example, consider two specific sequences PHHP and HPPH:



Both sequences have the same energy in the U-shape structure, which is an energy minimum for HPPH, but not for PHHP. In fact, the PHHP sequence does not have a unique ground state whereas the HPPH sequence does; the latter sequence is "protein-like" in that it has a unique and folded ground state structure.

Here, only HHHH, HPPH and HHPH have unique ground states; all other sequences are degenerate. The characteristic of sequences having a unique folded ground state is that the two ends of the polymer are both H sites, thus reducing the repulsive energy.

Longer polymers have many different folded states, only a subset of which may contain ground states for various sequences. For example, six linked sites on a square lattice possess three different fully folded states (asterisk indicates start point of the chain):



For the six-site system, there are $2^6 = 64$ distinct sequences in the HP model. Some of the sequences have a unique ground state structure as shown: HHPPHH has a U-

shape, PHHHPH has an S-shape and HHPHPH has a G-shape. Based on only the three folded configurations in the figure, it is not difficult to establish that:

- 9 sequences have a U-shaped ground state structure
- 6 have an S-shape
- 3 have a G-shape.

Thus, 18 out of the 64 sequences possess unique folded ground states, and all three of the compact structures can be a ground state for at least one sequence.

The larger the number of sequences having a particular structure for a ground state, the more *designable* that structure is. In the six-site system, the U-shaped structure is the most designable.

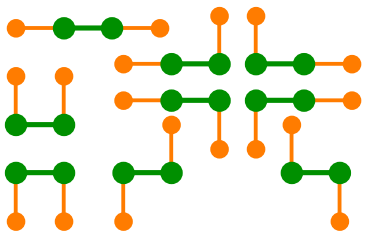
A structure is *encodable* if at least one sequence has that structure as its ground state configuration. The U, S and G-shaped structures above are all encodable: nine sequences encode for the U-shaped structure, six encode for the S-shape and three encode for the G-shape. In the four-site system on a lattice, only the U-shape is encodable and any sequence with an H at each end encodes for it.

A few general results from the study of small chains: about 2.1 - 2.4% of sequences on a two-dimensional square lattice have unique ground states (for chains with up to 18 segments; Chan and Dill, 1991).

Chain thermodynamics

The probability of finding a sequence in a given conformation involves the Boltzmann weight $\exp(-\Delta E/k_B T)$ and the degeneracy of the state g (the number of states with the same energy).

Consider a four-site HP system with the sequence HPPH, as displayed in panel (b) above. The figure does not represent correctly the degeneracy of the conformations in an ensemble. Suppose the two middle sites have a fixed location on a horizontal line, then the two end sites each can independently assume three orientations for a total of 3 × 3 = 9 conformations as shown: 2 U-shapes (up and down), 4 L-shapes, 2 zig-zags and 1 straight line:



All the shapes have energy $\Delta E = +2\varepsilon_H$ with respect to the U-shape, so the likelihood P_{gs} of finding the system in the ground state is

$$P_{gs} = 2 / [2 + 7 \cdot \exp(-2\varepsilon_H/k_B T)]. \quad (1)$$

The denominator is just the sum over Boltzmann weights including the degeneracy factors. At zero temperature, $P_{gs} = 1$ as expected, while at high temperature, P_{gs} approaches 2/9.

In the 4-site HPPH system, the mean interaction energy $\langle U \rangle$ of an ensemble of polymers is

$$\langle U \rangle = 14\varepsilon_H \exp(-2\beta\varepsilon_H) / [2 + 7 \cdot \exp(-2\beta\varepsilon_H)]. \quad (2)$$

if the interaction energy uses the ground state energy as a reference point. This follows because there are two states with energy zero, and seven states with energy $2\varepsilon_H$. As a function of temperature, $\langle U \rangle$ initially rises slowly from $T = 0$, then more rapidly as the population shifts to unfolded conformations, before slowly approaching its asymptotic value of (14/9) ε_H. This suggests that the model protein's specific heat will have a peak value in the temperature region where unfolding occurs. From Eq. (2), the specific heat C_V can be obtained analytically from the derivative of $\langle U \rangle$ with respect to temperature,

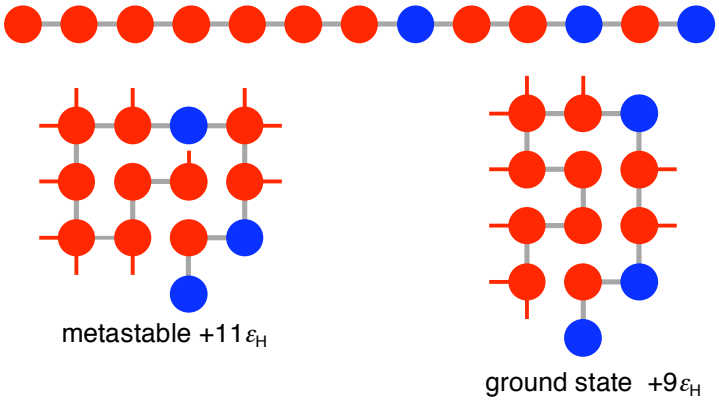
$$C_V = (\partial U / \partial T)_V. \quad (3)$$

Here, the lattice has a fixed spacing so the heat capacity is determined at constant volume, C_V, rather than constant pressure and has the form

$$C_V = 56\varepsilon_H^2 k_B \beta^2 \chi / (2 + 7\chi)^2. \quad \text{where } \chi = \exp(-2\beta\varepsilon_H) \quad (4)$$

The peak in C_V occurs at a temperature $k_B T$ of $0.685\varepsilon_H$, which is about 1/3 of the energy difference $2\varepsilon_H$ between the ground state and all other states.

On its way through a suite of conformations, the model protein may become "trapped" in a metastable configuration: a conformation that is not the true ground state of the sequence but has a lower energy than any conformation which can be reached from it by just a few moves of a bond or vertex. For example (from Chan and Dill, 1994)



The metastable configuration can only reach the ground state through a series of moves that take it through states with significantly higher energy.