24 Lattice Polymers

24.1 The Protein Folding Problem

To see how these rather abstract formulas apply to real-life systems, we'll explore next a simple model for protein folding. Protein folding is an incredibly complex process that in many ways defies simplistic explanations. Nevertheless, a few characteristic "driving forces" in protein folding can be identified:

- The **decrease in protein entropy** associated with folding into a single, well-defined conformation typically *opposes* folding. This observation offers a simple (in fact overly simplistic) explanation for why proteins tend to unfold at high temperatures. However, it should be born in mind that even a "folded" protein structure actually possesses some degree of conformational freedom since both the backbone and (especially) side chains are still able to flex and rearrange themselves locally, even in the folded state. This residual freedom helps somewhat to offset the entropy loss associated with folding.

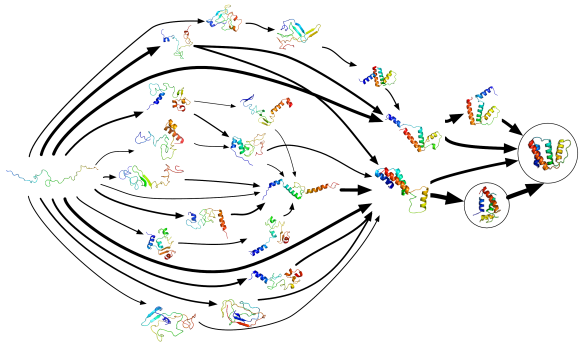
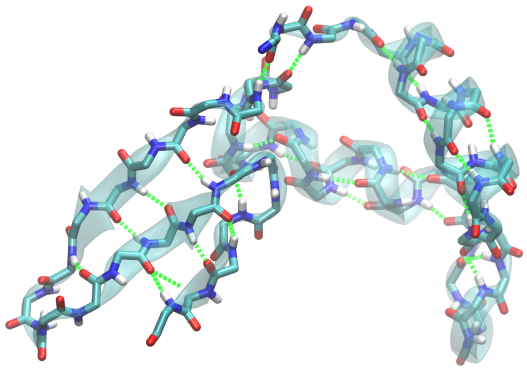
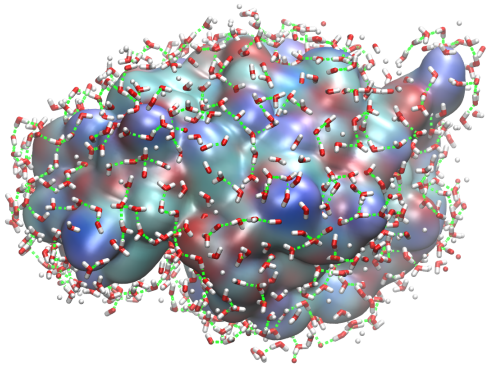


Image by Vincent Voelz, online at `https://commons.wikimedia.org/wiki/File:ACBP_MSM_from_Folding@home.tiff`. Related to publication Voelz et al. *J. Am. Chem. Soc.* 2012, 134, 12565-12577; `http://dx.doi.org/10.1021/ja302528z`

- **Hydrogen-bonding interactions** within the protein backbone and between polar sidechains stabilize the folded structure. In many cases, this contributes to a *negative enthalpy change* associated with protein folding. However, protein-folding also *decreases* the number of hydrogen-bonding interactions that the protein can form with the solvent (usually water), which to some extent counteracts the stabilization from protein-protein interactions.



- Protein folding usually **increases the entropy of the solvent** (usually water) since the folded state of the protein is usually more compact than the unfolded state; this leaves water molecules that would otherwise be occupied with protein-water interactions free to diffuse through the entire solvent volume, increasing the solvent entropy.



As you may have noticed, the various "driving forces" for protein folding are very much in competition with each other. Protein folding *decreases* the entropy of the protein but *increases* the entropy of the solvent; it *increases* the number of energetically favorable protein-protein hydrogen bonds but *decreases* the number of protein-water hydrogen bonds. And the effects we've described are only the tip of the iceberg! Other key factors that are even harder to quantify include the *flexibility* of the folded structure (which can substantially increase the entropy of the folded state) and, most importantly, the *specific interactions* enabled by the unique sequence of amino acids that compose the protein chain. Examples of sequence-specific features include salt bridges formed between Arg and Lys residues and the steric "kinks" introduced into the backbone by the presence of a Pro residue.

The complexity and interconnectedness of these effects leads to what is known as the *protein-folding problem*:

> The **Protein-Folding Problem** is a broad question in biophysical chemistry that asks *how the folded state of a protein can be predicted from its amino acid sequence* and *what mechanisms guide proteins to fold to their native structure.*

In the early days of biophysics, it was expected that relatively simple rules should be available to predict folded protein structures from the amino acid sequence, similar to the way a few simple hydrogen-bonding and base-stacking patterns account for the three-dimensional crystal structure of double-stranded DNA. As time passed, however, and as more crystal structures became available, it became gradually clearer that the protein-folding problem is much more complex than the parallel problem for DNA.

24.2 Lattice Proteins

In recent years, the *quality* of our computational models and the *quantity* of available computing power have together produced significant advances against this challenge. In what is becoming an impressive variety of cases, now, native protein structures *can* be predicted from only the amino acid sequence, and novel protein structures can even be designed from scratch. Nonetheless, the governing principles behind the problem remain complex since there are a huge number of degrees of freedom that must be simultaneously considered in analyzing the sequence of any given protein.

For exactly this reason, protein folding is very much a statistical problem and is ripe for analysis in terms of the canonical ensemble we've just described.

Since the full protein folding problem is much too complex for us to treat here, we'll consider a simplified problem know as the *lattice protein* model.

> A **Lattice Protein** model is a simplified description of the protein-folding problem, where the protein chain is modeled by a sequence of "beads" (representing amino acids) on a 2D or 3D grid (the "lattice").

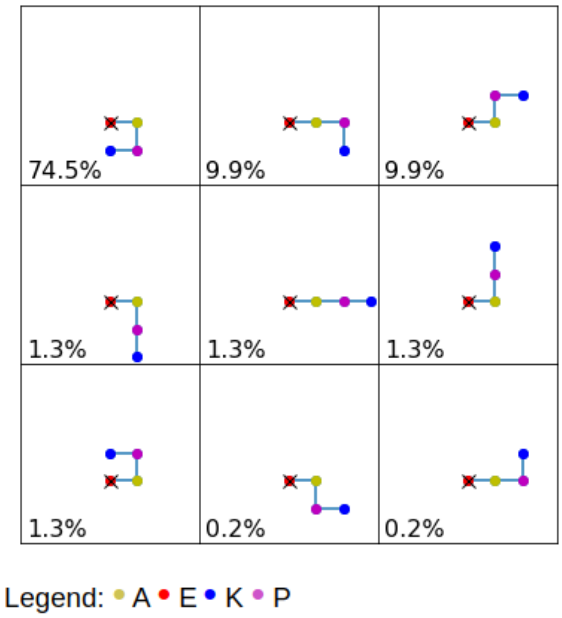
The use of a discrete lattice in this model (rather than the full, continuous, three-dimensional space available to real protein sequences) dramatically simplifies its analysis and in many cases allows us to enumerate conformational microstates explicitly.

To keep things simple, we'll think about protein conformations in only **two-dimensions** (2D), and we'll consider peptides made of only four amino acids with (dramatically!) simplified physical properties. Given a peptide sequence, we'll enumerate (computationally) all the possible conformations μ of the peptide on our 2D lattice, assign each conformation an energy ε_μ based on the amino-acid interactions in that structure, and then assign temperature-dependent probabilities in the canonical ensemble. Real proteins are, of course, much more complicated than the 2D models we'll study here, but this simplified model will at least give us a small taste of the kind of energetic and entropic forces the drive proteins to fold into stable, well-defined three-dimensional structures.

Specifically, we'll think about peptide sequences composed of four idealized amino acids:

- **Alanine** (Ala / A) residues will be treated as purely hydrophobic: A favorable energy shift $-\varepsilon_o$ will be assigned for each Ala-Ala contact in our peptide structure.
- **Lysine** (Lys / K) is a positively charged amino acid. We assign an unfavorable energy offset $+\varepsilon_o$ for each Lys-Lys contact and a favorable offset $-\varepsilon_o$ for every Lys-Glu contact (see next item).
- **Glutamate** (Glu / E) is a negatively charged amino acid. We assign an unfavorable energy offset $+\varepsilon_o$ for each Glu-Glu contact and a favorable offset $-\varepsilon_o$ for every Lys-Glu contact.
- **Proline** (Pro / P) is a sterically constrained amino acid due to the fact that its alkyl sidechain is chemically bonded to the amine nitrogen, forming a five-membered ring. This leads to somewhat complicated steric constraints on the roles Pro can serve in protein structures, but we'll simplify the situation in our 2D model to assign
  - A favorable energy shift of $-\varepsilon_o$ when the peptide chain makes a right-hand turn at a Pro residue
  - An unfavorable shift of $+\varepsilon_o$ when the peptide chain makes a left-hand turn at a Pro residue
  - No offset when the chain is straight at a Pro residue.

Consider, for example, the nine possible combinations of the tetra-peptide EAPK sketched in the diagram below. Here each dot in the diagram represents an amino acid, and the thin blue lines indicate the amide bonds connecting them.



The conformer in the top left corner has an assigned energy of

$$\varepsilon_\mu = -2\varepsilon_\mathrm{o} \tag{295}$$

since it is stabilized by both a Glu-Lys contact (salt bridge) and the favorable right-hand turn of the Pro residue. The next two conformers (top-center and top-right) are stabilized by favorable Pro conformations but not by any electrostatic contacts and hence are assigned slightly higher conformation energies of $\varepsilon_\mu = -\varepsilon_\mathrm{o}$. The three conformations in the middle row have energies of $\varepsilon_\mu = 0$ since there are no electrostatic or hydrophobic contacts, and the Pro residue is in the neutral (neither favored nor disfavored) straight conformation. The bottom-left conformer is also assigned $\varepsilon_\mu = 0$ since the favorable electrostatic interaction is counter-acted by an unfavorable Pro conformation. Finally, the bottom-center and bottom-right conformations are assigned the unstable value $\varepsilon_\mu = +\varepsilon_\mathrm{o}$ due to the disfavored left-turn Pro conformation and the lack of any stabilizing electrostatics.

From these conformation energies, we can calculate conformation *probabilities* for our peptide using the Canonical ensemble expressions of Eqs. (289) and (290). From the nine conformations in the figure above, we can calculate the partition function as

$$Z' = \sum_{\mu=1}^{9} e^{-\frac{(\varepsilon_\mu - \varepsilon_\mathrm{min})}{k_\mathrm{B}T}} \tag{296}$$

$$= e^0 + 2e^{-\beta\varepsilon_\mathrm{o}} + 4e^{-2\beta\varepsilon_\mathrm{o}} + 2e^{-3\beta\varepsilon_\mathrm{o}}. \tag{297}$$

Now suppose for concreteness that

$$\varepsilon_\mathrm{o} = 2 \cdot k_\mathrm{B} \cdot 300\,\mathrm{K}. \tag{298}$$

This gives a partition function at 300 K of

$$Z(300\,\mathrm{K}) = 1 + 2e^{-2} + 4e^{-4} + 2e^{-6} \approx 1.35. \tag{299}$$

Intuitively, this says that the system is *mostly* restricted to the lowest-energy conformation, with a small population in the higher-energy conformations. The populations can be calculated quantitatively using Eq. (292); for example, the population of the lowest-energy conformer is simply

$$p_0 = \frac{1}{Z'} \approx 74\%. \tag{300}$$

The remaining populations at $T = 300\,\mathrm{K}$ are reported in the lower-right corner of each figure.

24.3 LatticeProtein Simulation App

To get a more "hands-on" feel for these lattice polymer simulations, go to the LatticeProtein app at `https://nanohub.org/tools/LatticeProtein`. (NanoHUB is a science computing site administered in part by Purdue, in collaboration with other research universities; you can create a free account using your Purdue login credentials.) There you can design peptides with between 1 and 15 amino acids, chosen from the four options (A, E, K, and P) that we've described above. You'll use this app to complete your homework assignment.