2.3.4 The Random Energy Model (REM) for compact heteropolymers

Deep in the globular phase, the states of the compact polymer can be visualized as the collection of all maximally compact configurations. In a lattice version, these are self-avoiding walks that visit all sites, leaving no empty ones, and are referred to as Hamiltonian walks. The number of Hamiltonian walks also grows exponentially with the number of steps as $g^N$, but is much smaller than the number of self-avoiding walks (If there are $z$ neighbors for each lattice site, the number of non-self-avoiding walks that do not step back grows as $(z-1)^N$, and $(z-1)^N \gg g^N \gg g'N$.) For a homopolymer, all such configurations are equally likely, but in a heteropolymer the distinct interactions between different monomers leads to variations in energy. At low temperatures the lower energy states are preferred, and there can potentially be a phase transition to a specific (ground state) configuration. For biological molecules, there are non-specific attractive forces that tend to aggregate all monomers, whereas specific interactions select a particular (native) shape amongst the manifold of possible compact states.

While there is superficial resemblance between the freezing of water, and the polymer attaining a unique ground state, there is a crucial difference that needs to be addressed. The molecules of water (or any other fluid) are identical, and can be permuted in all possible ways within the specified solid structure. This is not the case for a heteropolymer in which a particular monomer along the chain has to be placed in a specific position in the ground state. (The analog for a fluid would be to number all- hitherto identical- molecules, and require a solid structure in which each molecule has to occupy a previously assigned position in the solid structure!) This difference is at the origin of the so-called Levinthal paradox which asks the question of how the amino-acids in a long protein can find the unique native state in reasonable time. With the number of possible states growing exponentially, say as $g'^N$ with $g' \approx 3$, even a short protein of length 100 has to find a native state amongst $3^{100} \approx 10^{47}$ possibilities. Even if each state could be sampled in $10^{-9}$ seconds, the required time to find the correct state would exceed the age of the universe.
As the simple lattice model already faces the above difficulty, we shall explore its potential resolution by considering all compact configurations for a multi-component heteropolymer. The energy of a configuration $\alpha$ is given by

$$E_\alpha = \sum_{\langle ab \rangle} V_{ab},$$

(2.58)

where the sum is over all non-polymeric nearest-neighbor pairs $\langle ab \rangle$, and $V_{ab}$ is the interaction energy assigned to a neighboring pair of monomers $a$ and $b$. The partition function is obtained from the sum

$$Z = \sum_\alpha e^{-\beta E_\alpha},$$

(2.59)

over the $g^N$ states. To make headway with this hard problem, we make the drastic approximation of assuming that the bond energies $V_{ab}$ are independent random variables. Subject to this assumption, the energies $E_\alpha$ are then also random variables. According to the central limit theorem, as long as the number of terms $N_B$ in Eq. (2.58) is large, the energies $E_\alpha$ are Gaussian distributed. The mean and variance of the distribution are given by

$$\langle E_\alpha \rangle = N_B \langle V_{ab} \rangle \equiv N \varepsilon_0, \quad \langle E_\alpha^2 \rangle_c = N_B \langle V_{ab}^2 \rangle_c \equiv N \sigma^2,$$

(2.60)

where noting that $N_B = (z - 2)N/2$ (of the $z$ nearest neighbors for each site of the lattice, two connect subsequent monomers along the chain; each bond is shared by two sites), we have folded the proportionality constants into the definitions of $\varepsilon_0$ and $\sigma^2$.

For large $N$, the probability distribution for the energy takes the Gaussian form

$$p(E) = \frac{1}{\sqrt{2\pi N \sigma^2}} \exp \left[ -\frac{(E - N \varepsilon_0)^2}{2N \sigma^2} \right].$$

(2.61)

Since the total number of states is $g^N$, the density of states is $\Omega(E) = g^N p(E)$, and the entropy of this random energy model (REM) is given by

$$S(E) = k_B \ln \Omega(E) = k_B \left[ N \ln g' - \frac{(E - N \varepsilon_0)^2}{2N \sigma^2} \right] - \frac{k_B}{2} \ln(2\pi N \sigma^2).$$

(2.62)
The last term is not extensive (proportional to $N$) and can be safely ignored.

According to Eq. (2.62), $S(E)$ is shaped like a parabola, but thermodynamic constraints imply that only a certain portion of this curve is physical. First, the temperature $T$ is obtained from the slope of the curve via $T^{-1} = dS/dE$. Positive temperatures require the entropy to increase with energy, and thus only the states with $E < N\varepsilon_0$ are physically accessible. Second, the entropy cannot be negative, and $S(E)$ should thus stick to zero for $E < E_c$, where $E_c$ is obtained as

$$S(E_c) = 0 \implies \frac{E_c}{N} = \varepsilon_0 - \sigma \sqrt{2 \ln g'}.$$

(Note the connection to the extreme value problem studied earlier: $E_c$ is also the mean value of the lowest of $g'^N$ energies randomly selected from $p(E)$.) The singularity of entropy at $E_c$ signifies a phase transition into a so-called ‘glass state’, at a temperature $T_c$ given by

$$\frac{1}{T_c} = \left. \frac{dS}{dE} \right|_{E_c} = -k_B \left( \frac{E_c - N\varepsilon_0}{N\sigma^2} \right) = k_B \frac{\sqrt{2 \ln g'}}{\sigma} \implies k_B T_c = \frac{\sigma}{\sqrt{2 \ln g'}}.$$

It is tempting to equate the freezing of the heteropolymer with the folding transition separating denatured and folded states of a protein. There is, however, a problem with such an interpretation, and the REM was initially introduced as a model of (so-called spin) glass. The reason is that the transition to ground state, obtained above via equilibrium thermodynamic arguments, is in fact not easily achieved due to kinetic constraints: As the temperature is lowered towards $T_c$, the number of accessible states decreases rapidly, but just above $T_c$, the system can still thermodynamically access a large number of states with energy close to $E_c$. It is unlikely that the lower energy configurations of the REM polymer in the vicinity of $E_c$ have much in common. To change its state, the polymer will then have to rearrange many of its monomers, running into high energy barriers in the process. Thus we expect that the kinetics of the REM polymer will slow down significantly on approaching $T_c$. The conclusion is that random sequences are not likely to fold, and that evolution must have converged upon special foldable sequences.
2.3.5 Designed REM for protein folding

The random energy model points to a so called *speed–stability paradox*: Stability of the low energy (native) state is achieved at temperatures below $T_c$, where the dynamics for arriving at such a state will be impossibly slow. This contradicts the observation that most proteins fold easily and in a short time. Of course proteins are not typical random heteropolymers, and are presumably “designed” through evolution for both function and ease of folding. Fortunately, we can mimic such “design” by a small modification of the REM; we only need to add to the continuum of random energy states, a single state with low energy ($E_n < E_c$) representing the native configuration.

With the added state at $E_n$, the system makes a transition to the native configuration (i.e. folds) at a temperature $T_f$, high enough that there are still many equivalent states to explore. The location of $T_f$, and the corresponding energy $E_f$, can be obtained by equating...
free energies or Boltzmann weights, and leads to the “tangent construction” whereby $T_f$ and $E_f$ are related to $E_n$ via
\[ \beta_f \equiv \frac{1}{k_B T_F} = \left. \frac{dS}{dE} \right|_{E_f} = \frac{S(E_f)/k_B}{E_f - E_n} = \frac{N \ln g' - (E_f - N\varepsilon_0)^2/(2N\sigma^2)}{E_f - E_n}. \] (2.65)

As depicted in the figure, the above result equates the slope of the tangent line from the point at $E_n$ computed in two different ways.

To justify the above result, note that in the canonical ensemble, the probability of finding the system in the native state is
\[ p_n = \frac{e^{-\beta E_n}}{Z(\beta)}, \quad \text{with} \quad Z(\beta) = e^{-\beta E_n} + \int dE \Omega(E) e^{-\beta E}. \] (2.66)

A phase transition in which $p_n$ changes discontinuously from zero to one occurs only in the thermodynamic limit of $N \to \infty$. For the system to have a well-behaved thermodynamic limit (in which case various thermodynamic identities involving entropy and temperature can be safely used), we must insist that the range of energies as well as $\ln \Omega(E)$ should be proportional to $N$; the former implies that $E_n \propto N$. If so, then at a particular value of $\beta$ a single value of energy $E$ completely dominates the partition function $Z(\beta)$. For the partition function in Eq. (2.66), the dominant value occurs for some $E \geq E_f$ for $\beta \leq \beta_f$, and for $E = E_n$ for $\beta > \beta_f$. The probability to find the system in its native state then jumps discontinuously from 0 to 1 at the point when the corresponding contributions to the partition function are equal, i.e. at
\[ e^{-\beta_f E_n} = \Omega(E_f)e^{-\beta_f E_f}, \] (2.67)

which after taking the logarithm leads to the tangent rule in Eq. (2.65).

We can eliminate $E_f$ in terms of $\beta_f$ by noting that $E = N\varepsilon_0 - N\sigma^2\beta$, and $\ln g' = (\beta_c\sigma)^2/2$. Using these expressions and defining a quantity $\beta_n = (E_n - N\varepsilon_0)/(N\sigma^2)$, the above equation reduces to
\[ \beta_f = \frac{\beta_n^2 - \beta_f^2}{-2\beta_f + 2\beta_n}. \] (2.68)

This can be rearranged as a quadratic equation with solution
\[ \beta_f = \beta_n - \sqrt{\beta_n^2 - \beta_c^2}. \] (2.69)

The ratio of the folding temperature to the REM freezing temperature is thus
\[ \frac{T_f}{T_c} = \frac{\beta_c}{\beta_f} = \frac{\beta_n}{\beta_c} + \sqrt{\left( \frac{\beta_n}{\beta_c} \right)^2 - 1}. \] (2.70)

Faster folding to the native state can be achieved at higher temperatures by increasing the energy difference between $E_n$ and $E_c$.

In the designed REM, the folding is still slow for temperatures less than $T_c$ as there are energy traps separated by high energy barriers that are hard to cross. However, in the a range of temperatures $T_c < T < T_n$ the assumption is that traps are not effective, with barriers that can be overcome by thermal fluctuations, while the native state is stable, once accessed, due to its lowered energy.
2.3.6 Lattice models

Our discussion of the phases of the designed REM was still largely based on elaborations of thermodynamic arguments. One difficulty is that within the REM the connection between different polymer structures, and the kinetic moves from one structure to another, are not specified. It is precisely the energy barriers encountered along such moves that determine the time scale for reaching equilibrium. These issues can be addressed by adding dynamical rules to the simple lattice models introduced earlier. The rules have to be local (such as pivoting a bond), but sufficient to allow the polymer to arrive from any state to another. In a Monte Carlo (MC) procedure, moves are attempted randomly, and accepted/rejected according to the Metropolis algorithm: If the energy change $\Delta E$ is negative the move is always accepted. However, moves that increase energy ($\Delta E > 0$) are still sometimes accepted with probability governed by the Boltzmann weight $e^{-\beta \Delta E}$. While the MC procedure is a far from realistic description of protein dynamics, it simple lattice model does face the issues of speed/stability, and has to overcome the Levinthal paradox.

In the 1990s many such simulations were performed, typically on a polymer with 27 sites, which can fold to a variety of compact configurations on a 3x3x3 cubic lattice. While the low energy states were compact the starting (unfolded) configuration could be chosen amongst the much larger number of swollen self-avoiding walks on the square lattice. Each of the 27 sites is labelled with an “amino-acid” $a_i$, while the energy of a configuration is $E(C) = \sum_{i<j} U(a_i, a_j) \Delta_{ij}(C)$, where $\Delta_{ij}(C)$ is adjacency matrix for configuration $C$, whose elements are 1 or 0 depending on whether or not the non-polymeric pair $(ij)$ is adjacent in configuration $C$, and $U(a_i, a_j)$ is the matrix of interactions between amino-acids. A random assignment of amino-acids to the sites of the polymers typically does not yield a foldable state; the corresponding polymer at low temperatures usually gets stuck in a swollen state at low temperature. One question is thus how to find a sequence that is foldable.

It is in fact possible to devise a different MC procedure that mimics ‘evolution’ to design a sequence that has low energy in a specified configuration. For a target “native” structure, characterized by an adjacency matrix $\delta_{ij}(n)$, the goal is to minimize $E_n = \sum U(a_i, a_j) \Delta_{ij}(n)$, over the set of all possible adjacency matrices. More precisely, one needs to minimize $Z_n \equiv (E_n - E_{ave})/\Sigma$, where $E_{ave}$ is an average energy of all structures, and $\Sigma$ is the corresponding variance, for a given sequence $\{a_i\}$. In principle, $Z_n$ has to be obtained by comparing all competing structures with the same number of bonds as the desired configuration, e.g. all compact structures. This is computationally difficult, and in an approximation analogous to REM, we can set $E_{ave} = N \varepsilon_0$ and $\Sigma = N \sigma^2$, as in Eq.(2.60), related to the mean and variance of $U(a_i, a_j)$, respectively. The “evolution” MC then proceeds as follows: start with a random sequence; attempt a move replacing the amino-acid at a site $i$ from $a_i$ to $a'_i$. Accept or reject the “mutation” according to the change $\Delta Z_n$, with Metropolis probabilities controlled by a “design” temperature $T_{des}$. A polymer designed by the above procedure, starting from a swollen initial state, does not fold at low temperatures, and typically is again caught in a trap. The same polymer simulated at intermediate temperatures may (or may not) fold. Thus, it is empirically possible to construct foldable sequences in the lattice model that to overcome the Levinthal paradox.
The average folding time obtained in such simulations has a non-monotonic dependence on temperature $T$; at high temperatures there are too many competing states, while at low temperatures the system is easily caught in a trap. A theoretical model for folding time, which reproduced these trends can be constructed, and is presented in the problem set.