FYSB12

Computer lab

November 13, 2017

Polymer statistics

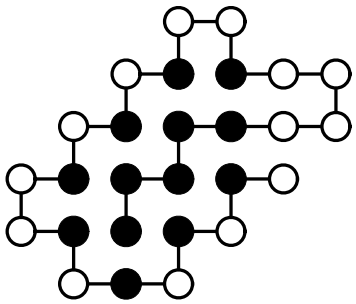
Anders Irbäck (e-mail: anders@thep.lu.se) and Daniel Nilsson (e-mail: daniel.nilsson@thep.lu.se)

INTRODUCTION

Proteins are polymer chains, made out of amino acids, that perform a wide range of functions in cells. Each protein has a unique, genetically determined sequence of amino acids, which can be written in a 20-letter alphabet, namely the 20 amino acids of natural proteins. One main class of proteins is that of globular proteins, which fold into compact, more or less well-defined shapes. A key force driving the folding is hydrophobicity. Hydrophobic, or apolar, amino acids tend to fold into the interior, whereas charged and polar amino acids end up on the surface of the protein. The number of possible arrangements of a protein chain grows exponentially with chain length, and is astronomically large even for a modestly sized protein with, say, 100 amino acids.

A minimal model that captures some basics of protein folding is the HP model,[1] where the protein chain is represented by a string of beads on a lattice. Each bead is of one of only two (rather than 20) types: either H (hydrophobic) or P (polar). Two beads cannot simultaneously share the same lattice site, and are said to be in contact if they are nearest neighbors on the lattice but not along the chain. The energy of a given configuration C is taken to be $E_C = -N_{\rm HH}\varepsilon$, where $N_{\rm HH}$ is the number of HH contacts and ε (> 0) is a parameter. This choice makes the formation of a core of H beads energetically favorable. Figure 1 shows an example of an HP sequence with 27 beads, in a state with $E_C = -13\varepsilon$. It can be shown that all the $>10^{10}$ other possible states of this HP chain have higher energy. The sequence may therefore be said to be protein-like, in the sense that it has a unique folded structure. Only a few per cent of all HP sequences have this property.

FIGURE 1: An HP chain with 27 beads in its unique minimum-energy state ($E_C = -13\varepsilon$). Filled and open circles represent H and P beads, respectively.



[1] K.F. Lau and K.A. Dill, *Macromolecules* **22**, 3986 (1989).

EXERCISES

The HP model can be studied on different lattices. For simplicity, throughout these exercises, a two-dimensional square lattice is used. The lattice spacing is denoted by a.

Below three exercises are described. The first two are mandatory and the third is optional. Try to solve exercises 1 and 2a before the lab. Data for exercises 2b and 2c (Table 2) and a simulation program for exercise 3 can be downloaded at http://home.thep.lu.se/∼anders/teaching/fysb12/.

1) High temperature

The probability of finding a given HP chain in a configuration C at temperature T is $P_C \propto \mathrm{e}^{-E_C/kT}$, where $E_C = -N_{\mathrm{HH}}\varepsilon$. At high temperature, all configurations become equally probable. The chain then behaves as a random walk, except for the self-avoidance condition (the chain must not cross itself).

Consider an ordinary random walk with N steps on the square lattice. Each step $\mathbf{s}_i$ has length $|\mathbf{s}_i| = a$ and is in one of four equally probable directions (up, down, right, left). The mean-square distance between the two end points is given by

$$r_{\mathrm{ee}}^2 = \left\langle (\mathbf{s}_1 + \ldots + \mathbf{s}_N)^2 \right\rangle = \sum_{i=1}^{N} \langle \mathbf{s}_i^2 \rangle + \sum_{i \neq j} \langle \mathbf{s}_i \cdot \mathbf{s}_j \rangle \tag{1}$$

where $\langle \cdot \rangle$ denotes an average over all possible realizations of the walk. In this equation, $\langle \mathbf{s}_i \cdot \mathbf{s}_j \rangle = 0$, because $\mathbf{s}_i$ and $\mathbf{s}_j$ are independent if $i \neq j$. It follows that $r_{\mathrm{ee}}^2 = Na^2$, and that therefore $r_{\mathrm{ee}} \propto N^{\nu}$ with $\nu = 1/2$.

After imposing the self-avoidance condition, it turns out that, for large N, r_{ee} still scales as N^{ν}, but with a different exponent ν. Table 1 shows data for r_{ee}^2 for a few different N for a self-avoiding walk. Plot the data in this table in log-log scale, along with the result for an ordinary random walk. Use the data to estimate the exponent ν for a self-avoiding walk (assuming that $r_{\mathrm{ee}} \propto N^{\nu}$).

2) Thermodynamic analysis based on the density of states

For an HP chain with a unique minimum-energy state, this single state will dominate at low temperatures, whereas all states are equally probable in the limit of high temperature. To find out how the transition between these two behaviors occurs, it is useful to analyze the heat capacity $C_V = d\langle E \rangle_T / dT$.

The thermal average of a general property f at temperature T can be written as

$$\langle f \rangle_T = \sum_C f_C P_C = \frac{\sum_C f_C \mathrm{e}^{-E_C/kT}}{\sum_C \mathrm{e}^{-E_C/kT}} \tag{2}$$

TABLE 1: Simulation data for r_{ee}^2 for a few different N, for a self-avoiding walk on the square latttice.

N	20	40	80	160
r_{ee}^2/a^2	66.7	193	549	1555

where f_C is the value of f in configuration C and the sums run over all possible configurations. The heat capacity C_V can, in principle, be obtained by using this equation to find $\langle E\rangle_T$ at different T, and then taking a numerical derivative. However, the computation can be simplified by making two observations. First, using equation 2, it can be shown that

$$C_V = \frac{d\langle E\rangle_T}{dT} = \frac{1}{kT^2}\left(\langle E^2\rangle_T - \langle E\rangle_T^2\right) \tag{3}$$

Second, for a property f that only depends on E, equation 2 can be rewritten as

$$\langle f\rangle_T = \frac{\sum_E f_E g_E \mathrm{e}^{-E/kT}}{\sum_E g_E \mathrm{e}^{-E/kT}} \tag{4}$$

where the sums are over energy levels rather than configurations and g_E counts the number of configurations with a given E. This expression contains much fewer terms than equation 2, but requires knowledge of the density of states, g_E. For short HP sequences ($\leq$27 beads), exact results for g_E are available. Table 2 lists g_E for the HP sequence shown in figure 1. Note that there are only 13 possible values of the energy, whereas the number of different configurations is $>10^{10}$.

In this exercise, you will use the known g_E (Table 2) to investigate how the behavior of this HP sequence depends on temperature. Proceed as follows.

(a) Starting from equation 2, show equation 3.

(b) Use equations 3 and 4 along with the data in Table 2 to compute C_V at different temperatures T. To avoid numerical instabilities, replace the $\mathrm{e}^{-E/kT}$ factors in equation 4 by $\mathrm{e}^{-(E-E_{\min})/kT}$, where $E_{\min} = -13\varepsilon$. In the calculations, set $\varepsilon = k = 1$ (E and T are then in units of ε and ε/k, respectively). Plot C_V against T for $0.1 < T < 1$ and estimate the temperature $T_{\max}$ at which C_V is maximal.

(c) The probability of finding the chain in its unique "native", or minimum-energy, state is given by

$$P_{\text{nat}} = \frac{\mathrm{e}^{-E_{\min}/kT}}{\sum_E g_E \mathrm{e}^{-E/kT}} \tag{5}$$

where $E_{\min} = -13\varepsilon$. Make a similar plot of P_{nat} against T, and compare the behavior of P_{nat} to that of C_V.

3) Thermodynamic Monte Carlo simulations

Determining the density of states by exact methods is feasible only for short chains. By using Monte Carlo methods, it is possible to study longer chains. In this exercise, this approach is illustrated using the same 27-bead chain (Figure 1) as an example.

TABLE 2: Density of states, g_E, for the 27-bead HP chain in figure 1. Each "state" corresponds to a pair of configurations related by reflection symmetry. One of all possible configurations (a straight line) has no symmetry-related partner. This "state" is therefore assigned a weight of 1/2.

E/ε	g_E
0	18 671 059 783.5
−1	15 687 265 041
−2	5 351 538 782
−3	1 222 946 058
−4	234 326 487
−5	40 339 545
−6	5 824 861
−7	710 407
−8	77 535
−9	9 046
−10	645
−11	86
−12	0
−13	1

The aim of a Monte Carlo simulation is to generate a sequence of configurations $C_1, \ldots, C_\tau$ distributed according to the desired probability distribution P_C. If the set of configurations is sufficiently large, then thermal averages can be estimated by averaging over these configurations, that is

$$\langle f \rangle_T = \sum_C f_C P_C \approx \frac{1}{\tau} \sum_{k=1}^{\tau} f_k$$

where f_k denotes the value of f in configuration C_k. The generated configurations are generally correlated and long simulations may be required in order to obtain a sufficient number of effectively independent configurations.

You will receive a ready-made Monte Carlo program for simulations of HP chains. Use this program to simulate the 27-bead sequence in figure 1 at the temperatures $T = 0.9\,T_{\max}$ and $T = 1.1\,T_{\max}$, where $T_{\max}$ is the maximum of the heat capacity (see exercise 2b).

During the course of a simulation, the program prints the Monte Carlo "time" and the energy to a file at regular intervals. Use the data in this file to compute C_V and P_{nat} at the two temperatures. Add these data points to the figures drawn in exercises 2b and 2c.