10. HP model for Protein Folding

Protein structures can be investigated using energy functions and trying to find the native structure by looking for the configuration with lowest energy. Often ignored in these computations are the facts that proteins obey laws of statistical mechanics, and that temperature is an important factor.

Statistical mechanics tells us that not every protein in a large group of them has the lowest energy. The energies are random but they obey certain statistical laws based on the Boltzmann distribution. Also it is known that proteins become disordered and unfolded at high temperature.

The Boltzmann distribution is a law of statistical mechanics related to random distributions of energies of a large number of molecules. The following general principle applies to many phenomena in statistical mechanics: *The most likely thing to happen is what is observed.* This tendency towards the most likely configuration is related to *entropy*. Entropy is a measure of randomness, and physical systems move towards higher entropy. The Boltzmann distribution is also important in the computational technique of simulated annealing. In this techniques parameters are varied randomly to search for the minimum of a function.

Probability is measured by counting the number of possibilities. A large number of molecules called an ensemble. Each molecule has a different energy. The probability of a certain energy is given by the number of molecules having that energy divided by the total number of molecules. Using combinatorial arguments we can explain the Boltzmann distribution

$$P(E) \sim A \exp(-\beta E) \tag{1}$$

where $P(E)$ is the probability that a molecule has energy E. Here β is a constant which is a function of temperature, and A is a constant chosen to assure that the total probability is 1. This distribution occurs in the situation where the energy is random, but the total energy is fixed. We see that the lowest energy is the most likely, but there is a probability that the molecule has higher energy.

10.1. **Random distribution of balls in boxes.**

To understand the Boltzmann distribution start with a simple combinatorial problem, balls tossed randomly into boxes. The balls can be thought of as molecules and the boxes as states. At first we ignore energy.

Combinatorial problem I: Given N balls placed randomly in M numbered boxes, what is the most probable distribution?

By distribution, we mean a list of the number of balls in each box. We list the number of ball in each box as $n_1, n_2, \ldots, n_M$.

This is really a counting problem. Think of numbering the balls and tossing them in sequence and assume that at each toss the ball has an equal chance of going in any box. This can be thought of in two different ways; either make a list of balls and the number of the box that it falls in, or make a list of the boxes and given the numbers of the balls in the box. Looking at it the first way, there are M^N possible outcomes of N tosses. We assume each is equally probable. Count the number of outcomes with a given distribution and divide by M^N to find its probability. Then find the distribution for the outcome of largest probability.

As an example, suppose $N = 7$ and $M = 3$. To distinguish the balls we think of them also as numbered, the number of the toss. One possible outcome with distribution 3,2,2 is

Box 1: balls 1, 3, 5
Box 2: balls 2, 7
Box 3: balls 4, 6

Another possibility giving the distribution 3,2,2 is

Box 1: balls 1, 3, 6
Box 2: balls 4, 5
Box 3: balls 2, 7

How many possibilities are there for the distribution 3,2,2? We can get all such distributions by taking a permutation of the numbers 1 to 7 and assigning the first three balls to box 1, the next two balls to box 2, and the last two to box 3. There are 7! permutations of 7 balls. Since we don't distinguish the order of balls in a box, we factor out the 3! permutations of balls in the first box, 2! in the second and 2! in the third. So there are

$$\frac{7!}{3!2!2!} = 210$$

possibilities for a 3,2,2 distribution.

If we take another distribution, say 1,1,5, then there are only

$$\frac{7!}{1!1!5!} = 42$$

possibilities, so 2,2,3 is a more probable distribution than 1,1,5. The balls are more evenly distributed for 2,2,3 since the numbers of balls in each box do not differ by more than 1.

So by a simple counting argument, the number of possible ways to toss N balls into M boxes with the distribution $n_1, n_2, \ldots, n_M$ is

$$\frac{N!}{n_1!n_2!\cdots n_M!}.$$

The number of possible outcomes is M^N, and so $1/M^N$ times this quantity is the probability of the distribution $n_1, \ldots, n_M$.

For the most probable distribution, the one with the greatest number of possibilities, the number of balls in each box is about the same. We can show that

$$\frac{1}{n_1!\, n_2! \cdots n_M!} \tag{2}$$

with

$$n_1 + n_2 + \cdots + n_M = N \tag{3}$$

is maximum when the balls when the values n_j are approximately equal.

If the distribution $n_1, n_2, \ldots, n_M$ maximizes (2) *under the constraint* (3) *then for any i and j, n_i and n_j differ by no more than* 1*, that is,*

$$n_i - n_j \leq 1.$$

Proof Consider a distribution with one less ball in box i and one more in box j. Since the original distribution was maximal, (2) is not increased. So replacing n_i and n_j by $n_i - 1$ and $n_j + 1$ respectively in (2) the result is less or equal. After cancellation, this gives

$$\frac{1}{(n_j+1)!(n_i-1)!} \leq \frac{1}{n_j!\, n_i!},$$

and so $n_i - n_j \leq 1$.

Thus in the most probable distribution, the balls are equally distributed in the boxes. The number of balls in any two boxes does not differ by more than 1.

10.2. **Random distribution of balls in boxes with energy.** To get a Boltzmann type distribution, consider a simplified situation where the energy is restricted to discrete numbers $1, \ldots, M$. Suppose that the number of each box indicates the energy of the ball in the box. Suppose also that the total amount of energy is fixed as a constant E_{tot}, so in addition to the constraint (3), we have the constraint

$$n_1 + 2\, n_2 + \cdots + j\, n_j + \cdots + M\, n_M = E_{\text{tot}}. \tag{4}$$

What is the most probable distribution in this situation? We will show that if the numbers n_j are very large, then the distribution is approximately exponential,

$$n_j \approx A \cdot B^j, \tag{5}$$

for constants A and B. Note that n_j/N represents the probability that for this distribution a molecule has energy j.

Note from (5) that

$$n_i/n_j \approx B^{i-j} \tag{6}$$

so that the quotient n_i/n_j depends only on the energy difference $i-j$.

To show that the most probable distribution is of the form (5), we show that the quotients n_{j-1}/n_j and n_j/n_{j+1} are equal.

If the distribution $n_1, n_2, \ldots, n_N$ maximizes (2) *under the constraints* (3) *and* (4) *and if the values n_j are large, then*

$$\frac{n_{j-1}}{n_j} \approx \frac{n_j}{n_{j+1}}. \tag{7}$$

for all j.

Proof. Let $i = j-1$ and $k = j+1$. Start with the maximal distribution and change it by considering a distribution with two less balls in box j and one more in box i and one more in box l. This does not change the total energy so constraint (4) still holds. Also we have not changed the number of balls, so (3) holds. We assumed that the original distribution maximized (2), so since since (2) decreases under this rearrangement, we have

$$\frac{1}{(n_i+1)!}\frac{1}{(n_j-2)!}\frac{1}{(n_k+1)!} \leq \frac{1}{n_i!}\frac{1}{n_j!}\frac{1}{n_k!},$$

and hence

$$\frac{n_j}{n_i+1}\frac{n_j-1}{n_k+1} \leq 1. \tag{8}$$

Likewise if by considering a distribution with two more balls in box j and one less in box i and one less ball in box k, get

$$1 \leq \frac{n_j+2}{n_i}\frac{n_j+1}{n_k}. \tag{9}$$

Letting $n_i \to \infty$ shows that

$$\frac{n_{i-1}}{n_i} = \frac{n_i}{n_{i+1}} \tag{10}$$

or

$$\frac{n_{i+1}}{n_i} = B \text{ for all } j \tag{11}$$

where B is a constant. It follows that

$$n_k = \frac{n_k}{n_{k-1}}\frac{n_{k-1}}{n_{k-2}}\cdots\frac{n_2}{n_1}\,n_1 = n_1\,B^{k-1} \tag{12}$$

and the distribution is *exponential* of the form $P(k) = n_k/N = A \cdot B^k$. Here $P(k)$ is the probability that for this distribution a ball is in box k. For the probabliities to add to 1,

$$A^{-1} = \sum_{k=1}^{M} B^k.$$

The expression on the right is called the *partition function.*

Using Calculus. The above argument is simple and does not require mathematics above counting and algebra. In most physics books the proof of Boltzmann's law is shown using calculus and the theory of Lagrange multipliers. If there is a very large number of balls it is best to think of

$$p_j = n_j/N$$

as a continuous variable with

$$\sum_{j=1}^{N} p_j = 1. \tag{13}$$

and use the *Stirling approximation*

$$\ln n! = n \ln n - n$$

for large n. So replace (2) by $\sum p_j \log p_j$, and the problem becomes to maximize

$$S = -\sum p_j \log p_j$$

under the constraints (13) and

$$\sum_{j=1}^{N} j\, p_j = E_{\text{tot}}. \tag{14}$$

This problem can be solved using *Lagrange multipliers*. Let

$$g = \sum_{j=1}^{N} p_j \qquad h = \sum_{j=1}^{N} j\, p_j,$$

then for $j = 1, \ldots, N$ solve

$$\frac{\partial S}{\partial p_j} - \alpha \frac{\partial g}{\partial p_j} - \beta \frac{\partial h}{\partial p_j} = 0$$

for constants α and β. Computing the partial derivatives, we get

$$-1 - \ln p_j - \alpha - \beta j = 0$$

$$p_j = e^{-1-\alpha} e^{-\beta j}$$

and again this shows the distribution is exponential, $p_j = A \cdot B^j$ where $B = e^{-\beta}$.

10.3. **The role of temperature.** In the case where energy is not in discrete units we can write $P(E) = A \exp(-\beta E)$, $0 < E < \infty$. Since the probability must integrate to 1, we have $A = \beta$, and we find, using integration by parts, that the average energy is given by

$$\text{Average energy} = \int_0^\infty P(E) E \, dE = \frac{1}{\beta}.$$

In the case of gases it is found that the average energy is proportional to the temperature, $1/\beta = kT$. This gives the Boltzmann distribution in its usual form (1).

10.4. **HP model.** The Boltzmann distribution together with the HP model for proteins is used to study thre phase transition from denatured to native state in proteins with decreasing temperature. The HP model uses a very simplified model of a protein. In this model, a protein is a chain, or discrete curve, of points labelled H or P depending on the particular amino acid. The points are assumed to be a distance of 1 apart. H denotes if the amino acid is hydrophylic and P denotes if it is hydrophobic. This is a typical classification of amino acids. Hydrophylic residues are expected to be on the outside of a protein and react well with water; hydrophobic residues do not interact with water and are usually found in the inside of a protein. A hydrophobic reside is usually not found next to a hydrophylic one.

References.

(1) *General Chemistry*, Linus Pauling
(2) *Molecular Driving Forces*, Dill and Bromberg