LECTURE 5. GENE REGULATION AND STATISTICAL MECHANICS

The Boltzmann distribution

This section follows Ken Dill's Molecular Driving Forces.

Let's see how the Boltzmann distribution arises from (i) counting multiplicity and (ii) maximizing entropy under an energy constraint.

Consider a system comprising N objects distributed across m categories, with n_i indistinguishable objects in each category $i = 1, \dots, m$. The multiplicity of a macrostate with a given set of n_i's is the number of permutations

$$W = \frac{N!}{n_1!\, n_2! \dots n_m!}.$$

Using Stirling's approximation $n! \sim n^n$ (the numerical factors cancel and the square root terms are negligible for large n_i's), this expression can be written in terms of the probabilities $p_i = n_i/N$,

$$\frac{1}{N} \log W = -\sum_{i=1}^{m} p_i \log p_i.$$

The RHS is known as the entropy S (per object). The entropy is the logarithm of the multiplicity because it must be an extensive property of the system: Given two subsystems, their multiplicities must multiply while their entropies must sum, and the logarithm is consistent with this requirement.

Entropy measures the disorder or the lack of information about a system. To see this, let's find the probability distribution that maximizes the entropy without any constraints other than the fact that $\sum p_i = 1$. This problem can be solved using the method of Lagrange multipliers, which states that the maximum of a function $f(x)$ under an equality constraint $g(x) = 0$ is the stationary point of the Lagrangian function $\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$, where λ is a scalar known as the Lagrange multiplier. For this problem, we get

$$\frac{\partial S}{\partial p_i} - \lambda = 0,$$

for all i. Thus, $-1 - \log p_i - \lambda = 0$ and $p_i = \exp(-1 - \lambda)$ is a constant. The distribution that maximizes the entropy without any constraints is flat.

What about maximizing the entropy under the constraint that the mean of the distribution is known? This problem is commonly encountered because the mean is often measurable whereas the full distribution is not. Let's assign a value ϵ_i to objects in category i. In this case, the average value $\langle \epsilon \rangle = \sum_i p_i \epsilon_i$ is known. For this problem, we have two constraints and therefore,

$$\frac{\partial S}{\partial p_i} - \lambda - \beta \epsilon_i = 0.$$

Thus, $p_i = \exp(-1 - \lambda - \beta \epsilon_i)$. The constant λ factor cancels when normalized, and we find that the distribution that maximizes the entropy given the mean is exponential

$$p_i = e^{-\beta \epsilon_i}.$$

This distribution is known as the Boltzmann distribution in statistical mechanics.

The Gillespie algorithm

Let's check that the Gillespie algorithm correctly implements a set of stochastic reactions with rate k_i. A straightforward implementation would be to discretize time into steps dt and check for each dt whether each reaction occurs or not. The time step dt must be small enough such that no more than one reaction occurs in a step. A more efficient implementation is to determine the joint probability $p(\tau, i)dt$ that the next reaction occurs during the interval τ and $\tau + dt$ and the reaction that occurs is i. The first part is equal to the probability for no reactions by time τ. Following the same reasoning as in Lecture 1, this probability is equal to $e^{-k\tau}$, where $k = \sum_i k_i$. The second part is equal to the probability that reaction i occurs during the interval τ and $\tau + dt$, which is $k_i dt$. Thus, $p(\tau, i) = e^{-k\tau} k_i$, which can be re-interpreted as $(ke^{-k\tau})(k_i/k) = p(\tau)p(i|\tau)$. The first part is the exponential waiting time distribution for a Poisson process with rate k, and the second part is the probability to pick reaction i among all reactions.