21.8: Specificity in Recognition and Binding

Specificity in Recognition

What determines the ability for a protein to recognize a specific target amongst many partners? To start, let's run a simple calculation. Take the case that a protein (transcription factor) has to recognize a string of n sequential nucleotides among a total of N bases in a dsDNA.

- Assume that each of the four bases (ATGC) is present with equal probability among the N bases, and that there are no enthalpic differences for binding to a particular base.
- Also, the recognition of a particular base is independent of the other bases in the sequence. (In practice this is a poor assumption).
- The probability of finding a particular n nucleotide sequence amongst all n nucleotide strings is

$$\left(\frac{1}{4}\right)^n \tag{21.8.1}$$

- For a particular n nucleotide sequence to be unique among a random sequence of N bases, we need

$$\left(\frac{1}{4}\right)^n \geq \frac{1}{N} \tag{21.8.2}$$

- Therefore we can say

$$n \geq \frac{\ln N}{\ln 4} \tag{21.8.3}$$

Example

For the case that you want to define a unique binding site among N = 65k base pairs:

- A sequence of n = ln (65000)/ln(4) ≈ 8 base pairs should statistically guarantee a unique binding site.
- n = 9 → 262 kbp

This example illustrates that simple statistical considerations and the diversity of base combinations can provide a certain level of specificity in binding, but that other considerations are important for high fidelity binding. These considerations include the energetics of binding, the presence of multiple binding motifs for a base, and base-sequence specific binding motifs.

Energetics of Binding

We also need to think about the strength of interaction. Let's assume that the transcription factor has a nonspecific binding interaction with DNA that is weak, but a strong interaction for the target sequence. We quantify these through:

ΔG_1: nonspecific binding

ΔG_2: specific binding

Next, let's consider the degeneracy of possible binding sites:

g_n: number of nonspecific binding sites = (N – n) or since N ≫ n: (N – n) ≈ N

g_s: number of sites that define the specific interaction: n

The probability of having a binding partner bound to a nonspecific sequence is

$$\begin{aligned} P_{\text{nonsp}} &= \frac{g_n e^{-\Delta G_1/kT}}{g_n e^{-\Delta G_1/kT} + g_s e^{-\Delta G_2/kT}} \\ &= \frac{(N-n)e^{-\Delta G_1/kT}}{(N-n)e^{-\Delta G_1/kT} + ne^{-\Delta G_2/kT}} \\ &= \frac{1}{1+\frac{n}{N}e^{-\Delta G/kT}} \end{aligned}$$

where $\Delta G = \Delta G_2 - \Delta G_1$. We do not want to have a high probability of nonspecific binding, so let's minimize P_{nonsp}. Solving for ΔG, and recognizing P_{nonsp} 1,

$$\Delta G \leq -k_B T \ln \left[\frac{N}{n P_{nonsp}} \right] \tag{21.8.4}$$

Suppose we want to have a probability of nonspecific binding to any region of DNA that is $P_{nonsp} \leq 1\%$. For $N = 10^6$ and $n = 10$, we find

$$\Delta G \approx -16 k_B T \qquad \text{or} \qquad -1.6 k_B T / nucleotide \tag{21.8.5}$$

for the probability that the partner being specifically bound with $P_{sp} > 99\%$.

Readings

1. G. Schreiber, G. Haran and H. X. Zhou, Fundamental aspects of protein−protein association kinetics, Chem. Rev. 109 (3), 839-860 (2009).
2. D. Shoup, G. Lipari and A. Szabo, Diffusion-controlled bimolecular reaction rates. The effect of rotational diffusion and orientation constraints, Biophys. J. 36 (3), 697-714 (1981).
3. D. Shoup and A. Szabo, Role of diffusion in ligand binding to macromolecules and cell-bound receptors, Biophys. J. 40 (1), 33-39 (1982).