8.592J–HST.452J: Statistical Physics in Biology

Assignment # **3**

Extreme Values

1. *Optimized Gaussians:* Let $x = \max\{r_1, r_2, \cdots, r_N\}$ be the largest of N independent, identically distributed Gaussian variables. Specifically, each r is distributed according to

$$p_1(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad \text{and generically set} \quad \overline{P}_1(x) \equiv \int_x^\infty dr p_1(r).$$

(a) Find the cumulative probability, $P_N(x)$ that the maximum is less than or equal to x.

(b) Show that in the limit of $N \gg 1$, x^*, the most likely value of x, can be obtained from either expression

$$\overline{P}_1(x^*) = \frac{1}{N}, \qquad \text{or} \qquad p_1'(x^*) + N p_1(x^*)^2 = 0,$$

and behaves as $x^* \simeq \sqrt{2\sigma^2 \ln N}$.

(c) By expanding the solution to part (a) around x^*, and expressing the first order expansion as an exponential (i.e. replacing $1 + \delta$ with e^δ), show that the probability distribution for x approaches a Gumbel form, and identify the corresponding parameters.

*****

2. *Homodimers versus heterodimers:* In Phys. Rev. Lett. **97**, 178101 (2006), Lukatsky, Zeldovich and Shakhnovich note that proteins are more likely to pair and interact as homodimers (two identical components) than heterodimers (with two distinct parts). They offer a statistical justification for this preference which is partly based on the charecteristics of extreme values. The simplified and analytically tractable model presented in this problem captures this aspect of the explanation.

We shall assume that the protein binding interfaces are circular rings of exactly N amino-acids. For a given ring, the amino-acids are selected randomly. A heterodimer is constructed by placing two such rings (a and b) in contact, and the resulting binding energy is

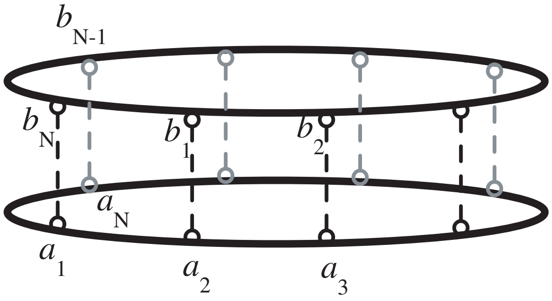
$$E_s^{a,b} = \sum_{i=1}^{N} V\left(a_i, b_{i+s}\right) \qquad , \qquad E^{a,b} = \min_s \left\{E_s^{a,b}\right\}.$$

Note that the two rings can be bound after relative shifts by $s = 1, 2, \cdots, N$, and the molecules rotate to achieve the location of minimal energy.

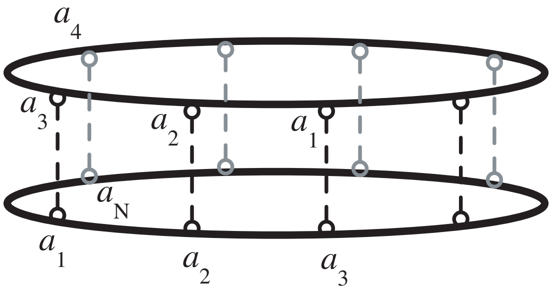
There are two ways to obtain a homodimer: The two sequences can be shifted and matched (not shown in the figure), in which case

$$E^{a,a} = \min_s \{E_s^{a,a}\} \quad \text{, with} \quad E_s^{a,a} = \sum_{i=1}^{N} V\left(a_i, a_{i+s}\right).$$

Heterodimer interface:



Homodimer interface:



However, since the rings are at the interface of a larger protein, such matching is generally not possible. The correct arrangement (as in the figure) is to rotate one of the two rings, and then join them, such that

$$E^{a,a_R} = \min_s \left\{E_s^{a,a_R}\right\} \quad , \text{with} \quad E_s^{a,a_R} = \sum_{i=1}^{N} V\left(a_{N-i}, a_{i+s}\right).$$

Throughout this problem assume that due to the addition of many pairwise interactions ($N \gg 1$), the energies E_s are Gaussian random variables, and that

$$\langle V(a,b)\rangle = 0, \qquad \text{while} \qquad \langle V(a,b)^2\rangle = \sigma^2.$$

(a) For the heteropolymers, find the mean $\left\langle E^{a,b}\right\rangle$ for $N \gg 1$, and comment on the form of the probability distribution for $E^{a,b}$.

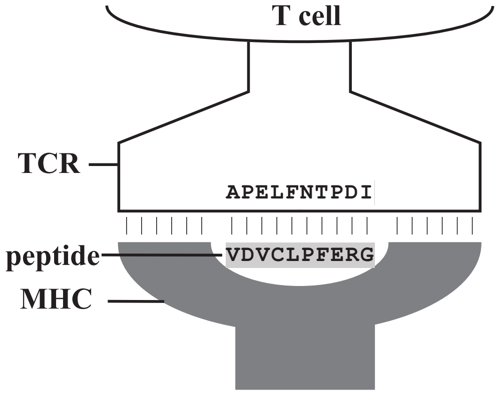
(b) For the un-rotated homodimers, find the probability distribution for $E^{a,a}$, and its mean value. (Hint: Note the number of distinct realization of s.)

(c) For the rotated homodimers, find the probability distribution for E^{a,a_R}, and its mean value. (Hint: Note the number of distinct interaction terms.)

(d) For randomly selected choices, which ensemble is likely to lead to (i) best binding; (ii) worst binding? If sequences are specifically designed to achieve optimal binding, is there any advantage to homodimers?

*****

3. *Thymic selection of T-cell receptors:* T cells are part of the adaptive immune system; their job is to examine short peptides cut from larger proteins and presented on the surface of cells in blood stream. The strength of binding between a receptor complex on the T-cell, and the peptide presented on another (major histocompatibility) complex, is used to determine whether the peptide comes from a self-protein or is part of a foreign pathogen protein. Pathogens are recognized when the variable T cell receptors (TCRs) bind strongly to foreign peptides; TCRs bind weakly to self-peptides and are thus self-tolerant. To ensure self-tolerance (thereby avoiding auto-immune response), the subset of T cells released to the blood stream is culled from a much larger candidate set in the thymus. In this problem, a simplified model of thymic selection of TCRs is mapped to an extreme value problem.



We shall assume that the relevant binding energy for discriminating between self and foreign peptides is due to an interface of N amino-acids from the peptide, and the TCR. The starting model thus resembles the previous problem, with

$$E(t,p)=\sum_{i=1}^{N}V(t_i,p_i),$$

where $t\equiv(t_1,t_2,\cdots,t_N)$ and $p\equiv(p_1,p_2,\cdots,p_N)$ indicate the sequences of amino-acids on the TCR and the peptide respectively. A given thymocyte (immature T cell) has some TCR sequence t; it moves around the thymus encountering cells presenting peptides from self-proteins. We shall assume that each thymocyte encounters M such peptides $\{p^{(\alpha)}\}$ for $\alpha=1,2,\cdots,M$. It is released into the blood stream only if two conditions are met:

$\star$ It must not bind any self-peptide too strongly. This condition, known as *negative selection* will be modeled by the requirement $E(t,p^{(\alpha)})>E_n$ for all α (negative energies correspond to stronger binding).

$\star$ It must bind at least one self-peptide moderately. This *positive selection* will be indicated by requiring $E_p>E(t,p^{(\beta)})>E_n$ for some β.

(a) Show that the above selection criteria are equivalent to $E_n<E_{\min}(t)<E_p$, where $E_{\min}(t)\equiv\min\{E(t,p^{(\alpha)})\}$ is the strongest binding energy.

(b) For a given TCR amino-acid t_i, let us set

$$\langle V(t_i, p_j)\rangle = \mathcal{E}(t_i), \qquad \text{and} \qquad \langle V(t_i, p_j)^2\rangle - \langle V(t_i, p_j)\rangle^2 = \mathcal{V}(t_i).$$

For large N, what is the probability distribution for $E(t, p)$ (for a given t encountering a random peptide).

(c) What is the probability distribution for $E_{\min}(t)$?

(d) Show that for $\ln M \propto N$, the distribution for $E_{\min}(t)$ is very narrow, centered around a value proportional to N, while its width does not grow with N.

*****

4. *Gapless alignment:* Consider a scheme for aligning DNA sequences in which a score $s = 1$ is assigned to a match, while $s = 0$ for a transversion (A↔G or T↔C) and $s = -\mu$ for a transition (e.g. A↔C). (Assume all four nucleotides occur with equal frequency.)

(a) Find the parameter $\lambda(\mu)$ that appears in the Gumbel distribution for the statistics of such random gapless alignments.

(b) Show that alignments are possible only for $\mu > \mu_c$, and plot λ as a function of μ.

*****