Computer practical: the optimal specificity of lymphocytes

Rob de Boer, Theoretical Biology, UU.

In the lecture we have defined the probability, P_i, of mounting an immune response to a single pathogen (epitope) as

$$P_i = 1 - (1-p)^R \quad \text{where} \quad R = R_0 P_s \quad \text{and} \quad P_s = (1-p)^S \tag{1}$$

define the post-selection repertoire, and the probability of surviving the self tolerance induction in the thymus, respectively. Whenever $p \ll 1$ this can be simplified into

$$P_i \simeq 1 - \mathrm{e}^{-pR_0P_s} \quad \text{where} \quad P_s \simeq \mathrm{e}^{-pS} \ . \tag{2}$$

This model has 3 parameters: S the number of self epitopes, R_0 the diversity of the pre-selection repertoire, and p the probability that a given antigen receptor binds the epitope of interest with sufficient avidity to activate the cell. We found that P_i is optimal when $p = 1/S$, giving that $P_s = 1/\mathrm{e}$, which resembles the secretary problem (`https://en.wikipedia.org/wiki/Secretary_problem`).

1. Probability of mounting a response, P_i

Use you favorite numerical environment (Mathematica, Python, Matlab, or R) to plot the probability of mounting an immune response, P_i, as a function the log of the specificity parameter, p. Use $R_0 = 10^9$ and $S = 10^5$ as default 'realistic' values for a human immune system. The aim of this exercise is to plot P_i for several values of S and R_0 to better understand why the immune system needs to be so diverse.

a. How do the results explained above depend on the values of S and R_0? How does each of these parameters affect the curve and the location of the optimum? What do you learn from this for the required diversity of the pre-selection repertoire? Phrase in your own words why the pre-selection repertoire should be so diverse.

b. For large R_0 the peak in the P_i curve is very wide. This could be an artifact of the fact that we only consider one pathogen here. Maybe the peak becomes more narrow when we consider the more natural problem of surviving a lot of pathogens. This is a simple extension of the model because the probability to survive n pathogens would just be P_i^n. Plot P_i^n as a function of $\log p$ to study how the likelihood to respond depends on the number of pathogens n. Does this narrow down the peak, i.e., is the range of 'good' immune systems becoming more narrow?

c. Since every pathogen consists of a large number of epitopes, one could also argue that the host is protected once it mounts an immune response to at least one of the epitopes. If there are n epitopes in a typical pathogen, the probability of 'at least one response' can be written as

$$P_i' = 1 - (1-P_i)^n = 1 - (1-p)^{Rn} \simeq 1 - \mathrm{e}^{npR} \ , \tag{3}$$

i.e., one minus the probability of no response to all n epitopes. Plot P_i' as a function of $\log p$ to study how the likelihood to respond depends on the number of episodes per pathogen n. What is now the optimum and does this narrow down the range of "good" immune systems?

d. Since most vertebrate species have a large genome, our estimate of about $S = 10^5$ self epitopes seems quite general. This suggests that the evolution of the adaptive immune system had to start with quite specific lymphocytes, and hence a large repertoire to be

functional. One of the smallest vertebrates is the fish species *Paedocypris*, which is known to have about $R = 37000$ T cells, and about 12000 self proteins [2]. With say 10 epitopes per protein the latter would indeed make $S = 10^5$ a reasonable estimate. What would be the probability, P_i, of an immune response to a foreign epitope for these fish?

2. Required diversity of the pre-selection repertoire, R_0

One criticism that one may have on the model optimizing the specificity of immune responses is that we have fixed the diversity of the pre-selection repertoire, R_0. One could argue that the loss of clonotypes due to negative selection can be compensated for by recombining novel receptors, i.e., by increasing R_0. One answer to this criticism would be that the R_0 defined in the models is the 'life time' pre-selection repertoire, i.e., the total number of unique receptors expected to be made. But one can also address this criticism by changing the question of an optimal immune response into the question 'How large would R_0 have to be if one were to demand an immune response to almost every foreign antigen?' In other words, 'How large an investment should a species make to achieve a protective functional repertoire?' One could argue that a repertoire becomes protective when every foreign antigen has a fair chance to be recognized by the repertoire as a whole. With a recognition probability of p per clone, this is achieved when $R \simeq 1/p$, as at that diversity an antigen is expected to be recognized by one clone. To address the question how diverse R_0 would have to be, we substitute $R = 1/p$ in $R = R_0(1-p)^S$ and solve for R_0,

$$\frac{1}{p} = R_0(1-p)^S \simeq R_0 \mathrm{e}^{-pS} \quad \text{or} \quad R_0 = \frac{1}{p(1-p)^S} \simeq \frac{\mathrm{e}^{pS}}{p} \ . \tag{4}$$

Use you favorite numerical environment (Mathematica, Python, Matlab, or R) to plot the required pre-selection repertoire, R_0, as a function the log of the specificity parameter, p.

a. What is the probability of mounting an immune response when $R = 1/p$? Hint use the approximation $P_i \simeq 1 - \mathrm{e}^{-pR}$.

b. How does the required diversity of the pre-selection repertoire depend on the specificity of its lymphocytes?

c. What do you think of the criticism that one can always generate novel clonotypes to fill up the functional repertoire?

d. How does the required R_0 depend on the number of self antigens?

e. Suppose every pathogen expresses several antigens (or epitopes), and that at least one response would be sufficient (i.e., consider $R = 1/(pn)$ where n is the number of epitopes per pathogen). How would that change the required investment in R_0?

f. We have discussed that some of the epitopes may overlap with self, and that this depends on the length of the peptides (strings) being used. For 9-mers this overlap was negligible, as there were about 10^7 unique 9-mers in self proteins, i.e., $c = \frac{10^7}{20^9} = \frac{1}{51200}$. For 5-mers with 2×10^6 unique 5-mers in the human self [1], the overlap is considerable, $c = \frac{2\times 10^6}{20^5} = 5/8$, i.e., about 50%. Can you add this overlap to the probability of mounting a response, P_i', and to the required investment in R_0? How would that affect the results?

3. Variation in the recognition probability, p

In our simple model all cells were considered to have the same specificity, p. If we allow for a ranges of specificities, the post-selection repertoire should also become specific because thymic selection would weed out the most crossreactive clones from the pre-selection repertoire (see Fig. 1). One can add this mechanism to the model by allowing for a range of specificities defined by a log normal distribution, with a mean μ and a standard deviation

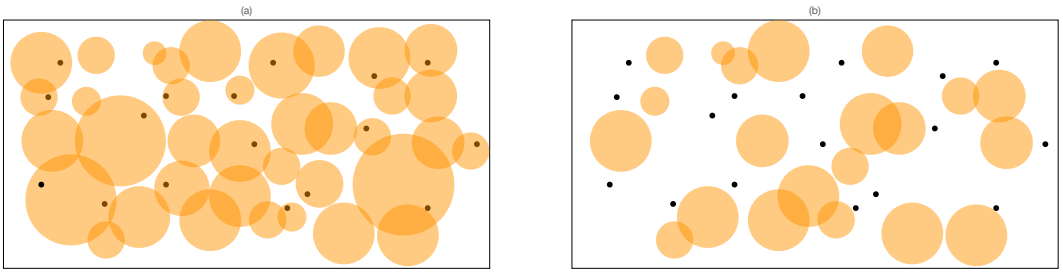

Figure 1: A cartoon of a pre-selection (a) and post-selection (b) repertoire in a shape space representation [3]. Clonotypes are depicted as orange discs representing the area in shape space that they cover. Self antigens are depicted as bullets. In Panel (b) we have deleted all clonotypes that cover at least one self antigen. This selects for the smaller orange discs.

σ (i.e., the log of the specificity, p, obeys a normal distribution),

$$D_0(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)} \text{ , for } x \leq 0 \text{ ,} \tag{5}$$

where $D_0(x)$ is a probability density, and $x = \log_{10} p$ defines a log specificity, i.e., $p = 10^x$. First, note that we are using a $\log_{10}$ for the specificity, instead of the conventional natural logarithm that is typical for a log normal distribution, because specificities are usually expressed as order of magnitudes (e.g., $10^{-6} < p < 10^{-5}$). This just scales the horizontal axis (with a factor $\mathrm{e} \simeq 2.73$). Second, $D_0(x)$ is probability density function having an area under the curve of one. To define the total number of clons, we therefore still need to multiply $D_0(x)$ with R_0 (i.e., $R_0(x) = R_0 D_0(x)$).

a. Plot the probability density function of Eq. (5) for various values of μ and for $\sigma = 1/2$ to note that for $\sigma = 1/2$ each repertoire contains a wide variation of antigen receptors, differing several orders of magnitude in their specificity.

b. Next use the same survival probability, $P_s \simeq \mathrm{e}^{-pS} = \mathrm{e}^{-10^x S}$ to define the remaining density of receptors in the post-selection repertoire as $D(x) = P_s(x) D_0(x)$, and plot $D(x)$ as a function of $\log p$ for varies values of μ.

c. How would you define the impact of negative selection?

d. How does this affect the results obtained with the model where p was the same for all clones.

e. Since clones with a lower specificity have a higher chance of surviving tolerance induction, but contribute with a lower probability to an immune response to a pathogen, we need to define contribution, $B(x)$, of every specificity to an immune responses (B stands fro breadth of the response). How would you write the equation for $B(x)$, and how would you define the breadth of the total immune response? If you know how to write these, plot them as a function of $\log_{10} p$ and $\log_{10} \mu$, respect

References

[1] **Burroughs, N. J., De Boer, R. J., and Kesmir, C.**, 2004. Discriminating self from nonself with short peptides from large proteomes. Immunogenetics. **56**:311–320.

[2] **Giorgetti, O. B., Shingate, P., O'Meara, C. P., Ravi, V., Pillai, N. E., Tay, B. H., Prasad, A., Iwanami, N., Tan, H. H., Schorpp, M., Venkatesh, B., and Boehm, T.**, 2021. Antigen receptor repertoires of one of the smallest known vertebrates. Sci. Adv. **7**:e0257016.

[3] **Perelson, A. S. and Oster, G. F.**, 1979. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. J. theor. Biol. **81**:645–670.

December 15, 2022