March 25, 2024

Computer Lab Exercise

Diversity of the Immune system

Handout for the *Immunobiology* lecture, at Utrecht University.

Rob J. de Boer

Objectives of this exercise:

1. See that immune systems are diverse because there are so many self antigens, and not because there are so many pathogens
2. See that the low number of MHC molecules per host optimizes the balance between positive and negative selection
3. Learn to use simple probabilistic models to ask evolutionary questions.

This handout and all files can be downloaded from the webpage of the diversity_practical.

Background

Diversity is a hallmark of the immune system. The repertoires of B cells and of $CD4^+$ and $CD8^+$ T cells each consist of more than 10^9 different clonotypes each characterized by a unique receptor (Qi *et al.*, 2014; Goncalves *et al.*, 2017). Each particular immune response is characterized by a large panel of different cytokines with –partly overlapping– functions, which is apparently functional for dealing with its cognate antigen. Each individual is characterized by a unique combination of MHC molecules that play an essential role in the selection of peptides that become presented to the cellular immune system. MHC loci are the most polymorphic genes known for vertebrates, i.e., for most loci many alleles have been identified within each population. However, each individual inherits only a limited number of MHC genes from its parents, and expresses about 10 different MHC molecules. We will here address the evolutionary questions why lymphocytes are so diverse within an individual, and why MHC molecules are diverse at the population level, and not diverse within an individual.

The consensus explanation for the enormous diversity of lymphocyte repertoires is the improved recognition of many different pathogens. The consensus explanation for the huge diversity (polymorphism) of MHC molecules in the population is that pathogens will not wipe out the entire population of hosts. There is no consensus why the MHC diversity within an individual is limited (although several authors argue that there would be too much negative deletion by self tolerance processes if the diversity of MHC molecules were higher).

Tolerance

We start with a simple toy model revealing some novel expectations for the relationships between the probability with which a lymphocyte binds an epitope, p, the number of self epitopes, S, and the pre-selection repertoire size, R_0 (De Boer and Perelson, 1993; Borghans *et al.*, 1999). Defining the lymphocyte binding probability, p, as the probability that a lymphocyte responds to a randomly chosen epitope, we have a definition that remains close to the conventional concept of the "precursor frequency" of an epitope. A typical viral epitope activates about one in 10^5 naive CD8$^+$ T cells (Blattman *et al.*, 2002; Su *et al.*, 2013; Kotturi *et al.*, 2008; Moon *et al.*, 2007; Tubo *et al.*, 2013; Jenkins *et al.*, 2010; Chang *et al.*, 2020).This means that the probability that a lymphocyte recognizes a randomly chosen epitope is about $p = 10^{-5}$. It is more difficult to estimate the number of self epitopes. For the peptides of nine amino acids (9-mers) that are used as epitopes by CD8$^+$ T cells, we have made an estimate by enumerating all unique 9-mers in the human genome (Burroughs *et al.*, 2004). Given that there are approximately 10^7 unique 9-mers in the human self, and that MHC molecules typically present about 1% of these, we would have an estimate of $S = 10^5$ self epitopes per T cell restricted to one particular MHC (Burroughs *et al.*, 2004). Fortunately, for the arguments presented here, the precise number of self epitopes turns out to be unimportant, we only need to know that it is large. The diversity of the pre-selection repertoire, R_0, is also a large number. The size of the functional T repertoire R in man is at least 10^9 different receptors (Qi *et al.*, 2014), and the diversity of the pre-selection repertoire is at least an order of magnitude higher because only a small fraction of the thymocytes expressing a functional $\alpha\beta$-TCR survive positive and negative selection in the thymus (see below).

Having these concepts at hand we write a simple mathematical model. The diversity of the functional repertoire R is determined by the chance that each clonotype in R_0 fails to recognize all self epitopes S, i.e.,

$$R = R_0(1-p)^S \,. \tag{1}$$

Similarly, the chance that an individual fails to respond to a foreign epitope is the probability that none of its clonotypes in the functional repertoire R recognize the epitope. Expressing one minus the chance of failure as the probability of mounting an immune response to a foreign epitope, we obtain

$$P_i = 1-(1-p)^R = 1-(1-p)^{R_0(1-p)^S} \,. \tag{2}$$

Using the R-script provided in the practical, P_i can be depicted as a function of the lymphocyte binding probability, p. Plotting p on a linear axis (Fig. 1a), and on a logarithmic axis (Fig. 1b), reveals that there is a very wide region of binding probabilities where the chance of mounting a successful immune response is close to one. If they are too cross-reactive, too many clonotypes are deleted by self tolerance processes, and the functional repertoire becomes too small (Fig. 1a and b), whatever the size of the pre-selection repertoire (Fig. 1c). If lymphocytes are too specific, i.e., at the left, epitopes remain unrecognized, but this can be compensated with a large pre-selection repertoire (Fig. 1c).

Because $(1-x)^n \simeq \mathrm{e}^{-xn}$ whenever $x \ll 1$ (see the footnote[1]), we can approximate this model by

$$R \simeq R_0\mathrm{e}^{-pS} \quad \text{and} \quad P_i \simeq 1-\mathrm{e}^{-pR} = 1-\mathrm{e}^{-pR_0\mathrm{e}^{-pS}} \,. \tag{3}$$

When plotted for the same parameters as those of Fig. 1 the approximation is indistinguishable from the original curve (not shown). The approximation allows us to compute the "optimal" value of P_i by taking the derivative $\partial_p P_i$ of Eq. (3) and solving $\partial_p P_i = 0$ to find[2] that that the maximum is at $\hat{p} = 1/S$. This optimum suggests that the lymphocyte binding probability is largely determined by the number of self epitopes the immune system has to be tolerant to. Thus, the binding probability is not determined by the recognition of pathogens, but by the demand to remain tolerant to a large number of self epitopes. Once lymphocytes are specific, the repertoire has to be sufficiently diverse to guarantee recognition of foreign epitopes (Fig. 1b).

For additional documentation you could read the paper by Borghans *et al.* (1999) who extend this model by allowing for self antigens that fail to induce tolerance, i.e., epitopes that are ignored. Healthy individuals do

[1] Why is $f = (1-x)^n \simeq \mathrm{e}^{-nx}$ when $x \to 0$? First, take the logarithm, $\ln[f] = n\ln[1-x]$. Second, using a linear approximation of $f(x)$ around a point $x = a$, i.e., $f(x)_{|x\to a} \simeq f(a) + (x-a)f'(x)_{|x=a}$, to write $\ln[f] \simeq n\ln[1] + xf'_{|x\to 0} = 0 + x\,\frac{-n}{1-x}|_{x\to 0} = -nx$. Hence $f \simeq \mathrm{e}^{-nx}$.

[2] Why is the derivative of $1-\mathrm{e}^{-a x \mathrm{e}^{-bx}}$ zero at $x = 1/b$? For this we have to apply the chain rule several times. First, consider $[\mathrm{e}^{-bx}]' = -b\mathrm{e}^{-bx}$. Next, use the product rule to see that $[-ax\mathrm{e}^{-bx}]' = -a\mathrm{e}^{-bx} + (-ax)(-b\mathrm{e}^{-bx}) = a(bx-1)\mathrm{e}^{-bx}$. Then we are ready to go for the whole expression $[1-\mathrm{e}^{-ax\mathrm{e}^{-bx}}]' = 0 - a(bx-1)\mathrm{e}^{-bx}\mathrm{e}^{-ax\mathrm{e}^{-bx}}$, which is zero when $x = 1/b$.

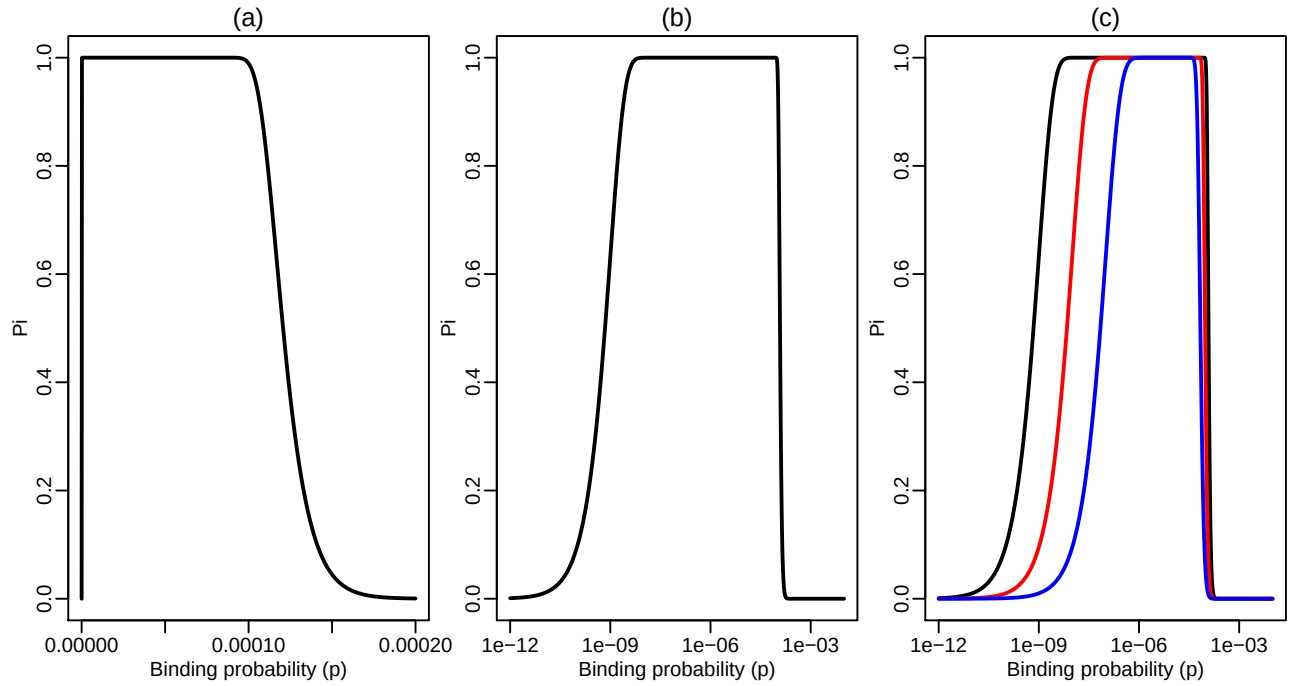

Figure 1: The probability of mounting an immune response P_i from Eq. (2) as a function of the binding probability p of the lymphocytes. Parameters $S = 10^5$ and $R_0 = 10^9$. Panel (a) and (b) depict p on a linear and a logarithmic scale, respectively. Panel (c) depicts the effect of decreasing the pre-selection repertoire size from $R_0 = 10^9$ (black), $R_0 = 10^8$ (red), to $R_0 = 10^7$ (blue). This reveals that immune systems with binding probabilities much larger than $1/S$ always perform poorly, whereas immune systems with very large pre-selection repertoires can afford their lymphocytes to have low binding probabilities. The black curve is the same in all panels.

harbor lymphocytes that can recognize self epitopes (Danke *et al.*, 2004; Su *et al.*, 2013; Yu *et al.*, 2015; Malhotra *et al.*, 2016). The optimum then shifts leftwards, i.e., towards a binding probability much smaller than $p = 1/S$. Thus, the $p = 1/S$ estimate (De Boer and Perelson, 1993) is an upper bound for the lymphocyte cross-reactivity: when the pre-selection repertoire is sufficiently large the immune system can allow for ignored self antigens when lymphocytes are more specific (Borghans *et al.*, 1999). Additionally, it is possible to allow for a distribution of binding probabilities in the pre-selection repertoire (**?**), because some TCRs appear to be more cross-reactive (i.e., have higher binding probabilities) than others (Lagattuta *et al.*, 2022; Textor *et al.*, 2023). This is interesting because negative selection is expected to weed out the most cross-reactive clones (Huseby *et al.*, 2003, 2005; Dai *et al.*, 2008; Chao *et al.*, 2005; Kosmrlj *et al.*, 2008), i.e., T cell repertoires also become more specific by selection in the thymus.

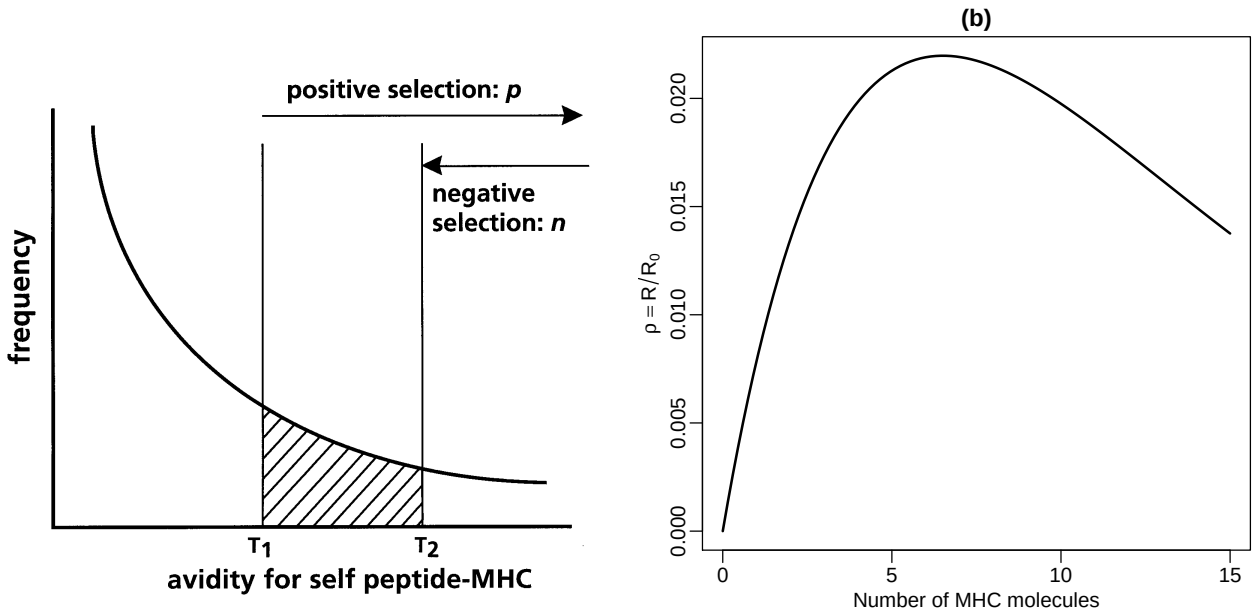

Figure 2: Positive and negative selection according to the avidity model (Janeway and Katz, 1984). The curve in (a) depicts the distribution of thymocyte avidities for self peptide–MHC complexes. In our model, the chance p to be positively selected by a single MHC type is the chance that the avidity between the thymocyte T cell receptor and any of the self peptide–MHC complexes exceeds threshold T_1. Thymocytes with avidities for any self peptide–MHC complexe exceeding the upper threshold T_2 are negatively selected (with chance n per MHC type). Panel (b) depicts the number of clones in the functional repertoire R, plotted as a fraction of the total pre-selection lymphocyte repertoire R_0 using Eq. (5). Parameters are: $p = 0.15, n = 0.14, \phi = 0.55$, and M is varied along the horizontal axis.

MHC diversity within the individual

Since increasing the individual diversity of MHC types would increase the presentation of pathogens to the immune system, one may wonder why the number of MHC genes is not much higher than it is. The argument that is typically invoked is that a higher MHC diversity within an individual would deplete T cell numbers by negative selection. However, this argument is incomplete, because more MHC diversity would also increase the number of clones in the T cell repertoire through positive selection (in order to be rescued in the thymus, lymphocytes need to recognize MHC-self peptide complexes with sufficient avidity). A high MHC diversity thus increases both the number of lymphocyte clones that are positively selected and the number of clones that are negatively selected. To calculate the net effect of these two opposing processes we need mathematical models (Nowak *et al.*, 1992; Borghans *et al.*, 2003), but these have come to opposing results. We will here derive an update of the model by Borghans *et al.* (2003). You can do a project to study why opposing results were obtained in the original papers.

The first step in positive selection is that a developing T cell should express a functional TCR. The β-chain of the TCR is tested first by binding with a pre-α-chain. Subsequently, T cells re-arrange their true α-chain on one of the two chromosomes. They can re-arrange the α-chain gene segments on the other chromosome when the first re-arrangement was not functional (this is somewhat sloppy and hence some T cells express two α-chains). Thus, due to the fact that only one third of the random re-arrangements leads to an in-frame receptor, and that T cells get a second change to re-arrange the TCRα gene segments on the other chromosome when the first re-arrangement is not functional, the probability of a functional α-chain re-arrangement is approximately $\phi = 0.33 + (1 - 0.33) \times 0.33 \simeq 0.55$.

Next consider an individual with M different MHC molecules and a pre-selection T lymphocyte repertoire consisting of R_0 different clones. Let p and n denote the probabilities that a clone is positively selected by a single MHC type, because its avidity to any self pMHC on that MHC is higher than a threshold T_1, or negatively selected because this avidity exceeds a higher threshold T_2, respectively (see Fig. 2a). By this definition, thymocytes can only be negatively selected by MHC molecules by which they are also positively

selected, i.e., $n < p$. Since T cell clones need to be positively selected by *at least one* of the MHC molecules, and avoid negative selection by *all* of the MHC molecules, the number of clones in the functional repertoire R can be expressed as

$$R = \phi R_0 \left((1-n)^M - (1-p)^M \right) , \tag{4}$$

where ϕR_0 is the repertoire of T cells that have successfully re-arranged their TCR. The functional repertoire R thus contains all TCR^+ T cell clones that are not negatively selected, minus the ones that fail to be positively selected by any of the M different MHC molecules of the host.

Experimental estimates for the parameters p and n of this model are difficult to obtain. The fraction of clones surviving, i.e.,

$$\rho = R/R_0 = \phi \left((1-n)^M - (1-p)^M \right) , \tag{5}$$

is known to be small in mice. Around 3% of the T cells produced in the thymus end up in the mature T cell repertoire (Goncalves *et al.*, 2017). The earlier modeling papers were based upon the estimate that least 50% of all positively selected T cells undergo negative selection (Van Meerwijk *et al.*, 1997; Merkenschlager *et al.*, 1997). More recent experiments argue that this is about 85% (Stritesky *et al.*, 2013). However, all of these estimates are based on the measured population sizes of double positive and single positive thymocytes, e.g., when a population halves after a selection step, one estimates that 50% of the cells have died. Since these population sizes also depend on the residence time at each stage, such steady state measurements are insufficient to estimate the fraction of clones surviving in between these stages.[3]

The best estimates therefore come from mathematical models estimating both the residence times and the population sizes (Sawicka *et al.*, 2014; Krueger *et al.*, 2017). By mathematical modeling thymocyte numbers in mice, Sawicka *et al.* (2014) propose a selection model for the fraction of cells surviving positive and negative selection

$$\rho = (1-p_1)(1-p_2)(1-p_3) = 0.0219 , \tag{6}$$

which is fortunately close to the 3% that was estimated previously (Goncalves *et al.*, 2017; Robert *et al.*, 2021). Here, $p_1 = 0.658$ is the fraction of cells dying from 'neglect' at the double positive stage (i.e., a survival of $\alpha = 1 - p_1 = 0.342$), $p_2 = 0.917$ is the fraction of cells dying when cells become single positive (i.e., a survival of $1 - p_2 = 0.083$), and $p_3 = 0.229$ is the fraction of cells dying at the single positive stage (i.e., a survival of $1 - p_3 = 0.771$). The total fraction of cells surviving negative selection therefore equals $(1-p_2)(1-p_3) = 0.064$, and hence the combined fraction of cells dying from negative selection is $\beta = 1 - 0.064 = 0.936$. Thus, $\alpha = 0.342$ is the fraction of cells that is positively selected, i.e., that have a functional TCR and have sufficient avidity for at least one of the MHC types, and $\beta = 0.936$ is the fraction of positively selected cells that are negatively selected. Their combination defines $\rho = \alpha(1-\beta) = 0.0219$.[4]

Taking into account that the inbred mice used in these experiments are homozygous and therefore express 3 types of class I MHC and 3 types of class II MHC molecules, i.e., a total of $M = 6$ MHCs, we need to solve p and n from Eq. (5) using $M = 6$, $\phi = 0.55$, $\alpha = 0.342$ and $\beta = 0.936$ (and hence $\rho = 0.0219$). Inspired by Eq. (4) we solve p from the probability of positive selection, $\alpha = \phi \left(1 - (1-p)^M \right)$,

$$\frac{\alpha}{\phi} = 1 - (1-p)^M \quad \text{or} \quad (1-p)^M = 1 - \frac{\alpha}{\phi} \quad \text{or} \quad 1 - p = \sqrt[M]{1 - \alpha/\phi} \quad \text{hence} \quad p = 1 - \sqrt[M]{1 - \alpha/\phi} \, . \tag{7}$$

It is tempting to solve n from the simple $1-\beta = (1-n)^M$, but this would be wrong because a cell that is positively selected on one particular MHC is not expected to be negatively selected on the other MHCs. We therefore

[3] To see that this is wrong consider the following cellular differentiation chain, $\mathrm{d}N/\mathrm{d}t = \sigma - (d_N + \delta_N)N$ and $\mathrm{d}M/\mathrm{d}t = f\delta_N N - (d_M + \delta_M)M$, where d_N and d_M are death rates, δ_N and δ_M are differentiation rates, f is the fraction of cells surviving the selection when N cells become M cells, and σ is the source of N cells. The expected number of cells in each population is defined by solving the the steady state $\mathrm{d}N/\mathrm{d}t = \mathrm{d}M/\mathrm{d}t = 0$, i.e., $\bar{N} = \sigma/(d_N + \delta_N)$ and $\bar{M} = f\delta_N \bar{N}/(d_M + \delta_M)$. The ratio of the two populations, $\bar{M}/\bar{N} = f\delta_N/(d_M + \delta_M)$ only defines the survival fraction, f, when $\delta_N = d_M + \delta_M$, i.e., when the rate at which the cells proceed through this cascade is the same for each cell type, $\delta_N = \delta_M$, and cells do not die, $d_M = 0$.

[4] Even this model is not completely correct because it is not strictly estimating the survival rates of clones as single positive cells are allowed to divide. This can be corrected by just considering the death and exit rates (Saccheri *et al.* work in progress), who estimate that $p_3 = 0.396$. Since the other cells types in model do not divide, the values of p_1 and p_2 remain the same. Hence we then obtain $\beta = 1 - (1-p_2)(1-p_3) = 0.95$ as the current best estimate, and since $\alpha = 1 - p_1 = 0.342$ stays the same, this corresponds to $\rho = 0.0171$, which fortunately remains close to the original estimate $\rho = 0.0219$.

solve n from Eq. (5), while substituting p from Eq. (7)

$$\frac{\rho}{\phi} = (1-n)^M - \left(1 - \frac{\alpha}{\phi}\right) \quad \text{or} \quad 1 + \frac{\rho-\alpha}{\phi} = (1-n)^M \quad \text{hence} \quad n = 1 - \sqrt[M]{1 + \frac{\rho-\alpha}{\phi}} \tag{8}$$

These two expressions for p and n enable us to estimate the probabilities of positive and negative selection on one MHC type from data providing the fraction of cells surviving positive thymic selection on all MHCs, α, and the fraction of cells surviving both positive and negative selection, ρ. A function in R solving p and n from α, $\rho = \alpha(1-\beta)$, and ϕ is explained and provided on the webpage numberMHC. For the parameters estimated above, i.e., $\alpha = 0.342$, $\beta = 0.936$, $\phi = 0.55$, and $M = 6$, we estimate the selection coefficients $p = 0.150$ and $n = 0.135$. The estimate of $p \simeq 0.15$ is not far from the observation that T cells bind about 20% of randomly selected MHC molecules (Dai *et al.*, 2008). Note that $p \simeq n$, which is strange and therefore interesting.

For these parameters the fraction $\rho = R/R_0$ is plotted as a function of the MHC diversity in Fig. 2b. We observe an optimum function that is maximal at around $M = 7$, which is close –but lower– than the true MHC diversity in outbred mice or humans (that should be ten to twenty in heterozygous individuals). Summarizing, the low MHC diversity within individuals is reasonably explained by the balance between positive and negative selection in the thymus.

Computer Lab Exercises

Today we will plot various functions using "R", which is a language with which one can easily do statistics, plot functions, and fit mathematical models to data. You should have installed R and RStudio on your laptop. The R-code for the functions that you need today is available on the webpage diversity. The easiest way to work with these function in RStudio is to copy-paste an R-chunk (gray box) from the website into the main window of Rstudio, highlight that (or put the cursor at the start), and then hit the `Run` button (or use the shortcut `control Enter`).

In the lecture we have defined the probability, P_i, of mounting an immune response to an epitope by Eq. (2). This model has 3 parameters: S the number of self epitopes, R_0 the diversity of the potential repertoire, and p the probability that a given antigen receptor binds the epitope of interest with sufficient avidity to activate the cell. The function `Pi()` from the webpage computes this probability, P_i, as a function of these 3 parameters, taking $R_0 = 10^9$, $S = 10^5$, and $p = 10^{-5}$ as default values. In the first line the functional repertoire R is calculated (see Eq. (1)), and in the last line the value of Eq. (2) is returned. One can now sketch P_i as a function the binding probability by calling the R-function `curve()`.

For instance, highlight the `Pi <- function(..)` function-definition, and hit the `Run` button to define the function in the R environment. Next, copy-paste the `print(c(Pi(), Pi(p=1e-4), Pi(r0=1e6)))` chunk, click at start of the line, and run it to print out P_i for different values of p and R_0. Next, copy-paste the `curve(Pi(p=x),..)` chunk and run it. This pops up a graphics window plotting P_i for its default parameters as a function of the binding probability, p. If you want to plot this function for another value of R_0, just use the up-arrow in the `Console` window to retrieve a previous command, and overwrite the R_0 parameter of the function, e.g., insert `r0=1e8` to obtain `curve(Pi(r0=1e8,p=x),from=1e-10,to=0.01,log="x")`.

Exercise 1 Probability of response

The aim of this exercise is to plot Eq. (2) 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 potential repertoire? Phrase in your own words why the potential 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. To plot P_i^n for say $n = 100$ pathogens, we just need to call `curve(Pi(p=x)^100,from=1e-10,to=0.01,log="x")` (see the webpage diversity). Modify your call to

`curve()` to study how P_i^n 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 $P_i' = 1 - (1 - P_i)^n$, i.e., one minus the probability of no response to all n epitopes. This expression is available as the function `PiN()` on the webpage diversity, and for the default value $n = 1$ this is identical to the `Pi()` function). 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, which seems an evolutionary challenge. 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 (Giorgetti *et al.*, 2021). With say 10 epitopes per protein the latter would indeed make $S = 10^5$ a reasonable estimate. What would be the probability of an immune response to a foreign epitope for these fish?

Exercise 2 What is the bottleneck during thymic selection?

Let us use the new estimates for positive and negative selection to define a simple quantitative scheme for the survival of thymocytes. We have estimated that a fraction $\alpha = 0.342 \simeq \frac{1}{3}$ survives positive selection. About half of this is due to cells successfully rearranging a functional $\alpha\beta$-TCR, i.e., $\phi = 0.55 \simeq \frac{1}{2}$. Thus the fraction, α', of $\alpha\beta$-TCR$^+$ clonotypes receiving a survival signal from the self pMHC that suffices to not die from neglect is solved from $\alpha = \phi\alpha' = \frac{1}{3} = \frac{1}{2}\alpha'$, meaning that a fraction $\alpha' = \frac{2}{3}$ receives a sufficient survival signal. For negative selection we estimated that $(1 - \beta) = 0.064 \simeq \frac{1}{20}$ of the clonotypes survives. Hence a simple rule of thumb for the combined selection process is

$$\rho \simeq \frac{1}{60} = \frac{1}{2} \times \frac{2}{3} \times \frac{1}{20} ,$$

which multiplies the probabilities to (1) successfully rearrange an $\alpha\beta$-TCR, (2) receive a sufficient signal from self pMHC, and (3) not a receive a strong signal from the self pMHC.

a. What is the major bottleneck during thymic selection?

b. Given that we estimated a probability of binding a single MHC expressing self peptides of $p \simeq 0.15$, what fraction of clonotypes in the post-selection repertoire, R, is expected to be restricted to more than one MHC? Hint, the probability of binding exactly i out of M MHCs is defined by a binomial expression, $\binom{M}{i} p^i (1-p)^{M-i}$, which for $i = 1$, $M = 6$, and $p = 0.15$ can be written in R as `dbinom(1,6,0.15)`.

c. What would be an estimate for R_0 if only $\frac{2}{3} \times \frac{1}{20} = \frac{1}{30}$ of the successful $\alpha\beta$ rearrangements survive thymic selection, and there are about $R = 10^9$ T cell clonotypes (Qi *et al.*, 2014)?

Exercise 3 Required diversity of R_0

In addition to maximizing the probability of mounting an immune response by optimizing the binding probability, p, one can also compute the size of the pre-selection repertoire required for having a sufficiently complete functional repertoire (for any given value of p). In other words, one can ask the question: How large an investment should a species make to achieve a protective functional repertoire (De Boer and Perelson, 1993)? For this we only need Eq. (1) defining the diversity of the tolerized 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 binding probability of p per clone, this is achieved when $R \simeq 1/p$, as at that diversity each antigen is expected to be recognized by exactly one clone. To address the question how diverse R_0 would have to be, we substitute $R = 1/p$ in Eq. (1) and solve for R_0,

$$\frac{1}{p} = R_0(1-p)^S \simeq R_0 \mathrm{e}^{-pS} \quad \text{or} \quad R_0 = \frac{1}{p(1-p)^S} \simeq \frac{\mathrm{e}^{pS}}{p} . \tag{9}$$

The latter expression is provided as the function `R0()` on the webpage diversity.

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 binding probability of its lymphocytes?

c. How does the required R_0 depend on the number of self antigens?
d. We can use these new insights to reconsider the very small fish *Paedocypris* having $R = 37000$ T cells (Giorgetti *et al.*, 2021). Above we assumed that this fish had an optimal binding probability of $p = 10^{-5}$. Would the fish do better with a higher binding probability?

Project

If you enjoyed this computer practical, you could decide to embark on a somewhat larger project, by comparing the model of Eq. (5) with the two previous publications (Nowak *et al.*, 1992; Borghans *et al.*, 2003). You should familiarize yourself with the estimation of p and n from the fractions of cells surviving positive selection, α in Eq. (7), and those dying from negative selection, β in Eq. (8). To this end we ask you in Exercise 4 to reproduce Fig. 2b and play with its parameters to study the location of the optimum as a function of the parameters. Next, embark on studying the models of the two previous publications (Nowak *et al.*, 1992; Borghans *et al.*, 2003) to understand what went wrong in these papers, and to convince yourself that Eq. (5) with Fig. 2b is indeed the most correct model. Read some of the papers that we cite, convince yourself fact that a large part of the positive selection in the thymus is just for productive rearrangements, and not for binding MHC or self antigens, find more papers, address (some of) the questions given below, and try to add some original results. The R-chunks on the webpage numberMHC can be used to work on the following exercises.

Exercise 4 General optimal #MHC

The curve in Fig. 2b depends on two parameters, p and n, that were estimated from the Sawicka *et al.* (2014) model, with the realization that $\phi = 0.55$ (Krueger *et al.*, 2017). The equations for solving p from α, Eq. (7), and n from β, Eq. (8), are available as the R-function `Rho()` on the webpage numberMHC.
a. The survival probability of positive selection, $\alpha = 0.342$ is defined as $\alpha = \phi(1-(1-p)^M) = \phi\alpha'$, where $\phi = 0.55$ and $\alpha' = 1-(1-p)^M$ is the probability of binding at least one MHC molecule. What is the relative contribution to positive selection of this requirement, α', of binding at least one MHC molecule in a mouse with $M = 6$ MHC molecules? .
b. What is the effect of changing the fraction of functional receptors, ϕ, on the optimal number of MHC types?
c. The current estimate of negative selection, $\beta = 0.936$, is not precise because Sawicka *et al.* (2014) allow cells to divide, which should not be included in the survival probability of a particular clone. In the footnote on page 5 we argue that $\beta = 0.95$ (and hence $\rho = 0.0171$). How sensitive are the results to these exact values?
d. What is your favorite explanation for the fact that the diversity of MHC molecules per host is low, while their degree of polymorphism of the population is high?

Exercise 5 Nowak's optimal #MHC

Nowak *et al.* (1992) also addressed the question of the optimum number of MHC molecules within an individual, using a similar approach but a different mathematical model. They write that a clone should become positively selected on at least one MHC molecule, with (for $\phi = 1$) the same probability, $\alpha = 1-(1-p)^M$, as we used in Eq. (4), but they define n^* as the *conditional* probability that a positively selected clone is negatively selected by a MHC molecule. Hence, $(1-n^*)^M$ is the probability that a positive selected clone is not negatively selected by any MHC molecule. Thus, according to their model the fraction of T-cell clones surviving selection in the thymus is

$$\rho = (1-(1-p)^M)(1-n^*)^M = \alpha(1-n^*)^M \,. \tag{10}$$

Note that this model can easily be extended by integrating the probability of successful recombination, ϕ, into α. Since the probability of positive selection is the same, one can calculate p from Eq. (7), which for the original $\phi = 1$ gives $p = 0.067$. To estimate the parameter n^* of this model we observe that

$$\rho = \alpha(1-\beta) = \alpha(1-n^*)^M \quad \text{or} \quad 1-\beta = (1-n^*)^M \quad \text{or} \quad \sqrt[M]{1-\beta} = 1-n^* \quad \text{hence} \quad n^* = 1-\sqrt[M]{1-\beta}\,, \tag{11}$$

which for $\beta = 0.936$ gives $n^* = 0.367$. Eq. (11) is available as the R-function `Nowak()` on the webpage numberMHC.
a. What is wrong with multiplying the probability of positive selection, α, with a term, $(1-n^*)^M$, defining the probability of not becoming negatively selected on any MHC? Hint: on how many MHC types is a clone expected to become positively selected? .

b. We find that $n^* > p$. Is that fair? Above we argued that $n < p$.
c. Study how the number of clones in the functional repertoire R, plotted as a fraction of the total initial lymphocyte repertoire R_0, i.e., ρ as defined by Eq. (10), depends on the number of MHC molecules, M. Do this for the same parameters as used in Fig. 2b.

Exercise 6 Borghans's optimal #MHC

The model derived in Eq. (4) is a simple extension of a model originally proposed by Borghans *et al.* (2003). We have just added the probability of successful recombination to positive selection, ϕ, which multiplies the whole equation, i.e., they used the model of Eq. (4) with $\phi = 1$. Since they obtained a much higher optimum (around $M = 150$), they argued that the number of MHC types per host is below the optimum, and that the earlier lower optimum of the Nowak *et al.* (1992) model (who also used $\phi = 1$) was due to the conceptual mistake studied in the previous exercise.
a. Explain why their optimum is much higher. Is there something wrong with this model?

References

Blattman, J. N., Antia, R., Sourdive, D. J., Wang, X., Kaech, S. M., Murali-Krishna, K., Altman, J. D., and Ahmed, R., 2002. Estimating the precursor frequency of naive antigen-specific CD8 T cells. J. Exp. Med. **195**:657–664.

Borghans, J. A. M., Noest, A. J., and De Boer, R. J., 1999. How specific should immunological memory be? J. Immunol. **163**:569–575.

Borghans, J. A. M., Noest, A. J., and De Boer, R. J., 2003. Thymic selection does not limit the individual MHC diversity. Eur. J. Immunol. **33**:3353–3358.

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.

Chang, Y. M., Wieland, A., Li, Z. R., Im, S. J., McGuire, D. J., Kissick, H. T., Antia, R., and Ahmed, R., 2020. T Cell Receptor Diversity and Lineage Relationship between Virus-Specific CD8 T Cell Subsets during Chronic Lymphocytic Choriomeningitis Virus Infection. J. Virol. **94**.

Chao, D. L., Davenport, M. P., Forrest, S., and Perelson, A. S., 2005. The effects of thymic selection on the range of T cell cross-reactivity. Eur. J. Immunol. **35**:3452–3459.

Dai, S., Huseby, E. S., Rubtsova, K., Scott-Browne, J., Crawford, F., Macdonald, W. A., Marrack, P., and Kappler, J. W., 2008. Crossreactive T Cells spotlight the germline rules for alphabeta T cell-receptor interactions with MHC molecules. Immunity **28**:324–334.

Danke, N. A., Koelle, D. M., Yee, C., Beheray, S., and Kwok, W. W., 2004. Autoreactive T cells in healthy individuals. J. Immunol. **172**:5967–5972.

De Boer, R. J. and Perelson, A. S., 1993. How diverse should the immune system be? Proc. R. Soc. Lond., B, Biol. Sci. **252**:171–175.

Garcia, K. C., Adams, J. J., Feng, D., and Ely, L. K., 2009. The molecular basis of TCR germline bias for MHC is surprisingly simple. Nat. Immunol. **10**:143–147.

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.

Goncalves, P., Ferrarini, M., Molina-Paris, C., Lythe, G., Vasseur, F., Lim, A., Rocha, B., and Azogui, O., 2017. A new mechanism shapes the naive CD8$^+$ T cell repertoire: the selection for full diversity. Mol. Immunol. **85**:66–80.

Huseby, E. S., Crawford, F., White, J., Kappler, J., and Marrack, P., 2003. Negative selection imparts peptide specificity to the mature T cell repertoire. Proc. Natl. Acad. Sci. U.S.A. **100**:11565–11570.

Huseby, E. S., White, J., Crawford, F., Vass, T., Becker, D., Pinilla, C., Marrack, P., and Kappler, J. W., 2005. How the T cell repertoire becomes peptide and MHC specific. Cell **122**:247–260.

Janeway, Jr, C. A. and Katz, M. E., 1984. Self Ia-recognizing T cells undergo an ordered series of interactions with Ia-bearing substrate cells of defined function during their development: a model. Surv. Immunol. Res. **3**:45–54.

Jenkins, M. K., Chu, H. H., McLachlan, J. B., and Moon, J. J., 2010. On the composition of the preimmune repertoire of T cells specific for peptide-major histocompatibility complex ligands. Annu. Rev. Immunol. **28**:275–294.

Kosmrlj, A., Jha, A. K., Huseby, E. S., Kardar, M., and Chakraborty, A. K., 2008. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. Proc. Natl. Acad. Sci. U.S.A. **105**:16671–16676.

Kotturi, M. F., Scott, I., Wolfe, T., Peters, B., Sidney, J., Cheroutre, H., Von Herrath, M. G., Buchmeier, M. J., Grey, H., and Sette, A., 2008. Naive precursor frequencies and MHC binding rather than the degree of epitope diversity shape CD8$^+$ T cell immunodominance. J. Immunol. **181**:2124–2133.

Krueger, A., Zietara, N., and Lyszkiewicz, M., 2017. T Cell Development by the Numbers. Trends Immunol. **38**:128–139.

Lagattuta, K. A., Kang, J. B., Nathan, A., Pauken, K. E., Jonsson, A. H., Rao, D. A., Sharpe, A. H., Ishigaki, K., and Raychaudhuri, S., 2022. Repertoire analyses reveal T cell antigen receptor sequence features that influence T cell fate. Nat. Immunol. **23**:446–457.

Malhotra, D., Linehan, J. L., Dileepan, T., Lee, Y. J., Purtha, W. E., Lu, J. V., Nelson, R. W., Fife, B. T., Orr, H. T., Anderson, M. S., Hogquist, K. A., and Jenkins, M. K., 2016. Tolerance is established in polyclonal $CD4^+$ T cells by distinct mechanisms, according to self-peptide expression patterns. Nat. Immunol. **17**:187–195.

Merkenschlager, M., Graf, D., Lovatt, M., Bommhardt, U., Zamoyska, R., and Fisher, A. G., 1997. How many thymocytes audition for selection? J. Exp. Med. **186**:1149–1158.

Moon, J. J., Chu, H. H., Pepper, M., McSorley, S. J., Jameson, S. C., Kedl, R. M., and Jenkins, M. K., 2007. Naive $CD4(^+)$ T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. Immunity **27**:203–213.

Muller, V., De Boer, R. J., Bonhoeffer, S., and Szathmary, E., 2018. An evolutionary perspective on the systems of adaptive immunity. Biol. Rev. Camb. Philos. Soc. **93**:505–528.

Nowak, M. A., Tarczy-Hornoch, K., and Austyn, J. M., 1992. The optimal number of major histocompatibility complex molecules in an individual. Proc. Natl. Acad. Sci. U.S.A. **89**:10896–10899.

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J. Y., Olshen, R. A., Weyand, C. M., Boyd, S. D., and Goronzy, J. J., 2014. Diversity and clonal selection in the human T-cell repertoire. Proc. Natl. Acad. Sci. U.S.A. **111**:13139–13144.

Robert, P., Kunze-Schumacher, H., Greiff, V., and Krueger, A., 2021. Modelling the dynamics of t-cell development in the thymus. Entropy .

Sawicka, M., Stritesky, G. L., Reynolds, J., Abourashchi, N., Lythe, G., Molina-París, C., and Hogquist, K. A., 2014. From pre-DP, post-DP, SP4, and SP8 Thymocyte Cell Counts to a Dynamical Model of Cortical and Medullary Selection. Front. Immunol. **5**:19.

Stritesky, G. L., Xing, Y., Erickson, J. R., Kalekar, L. A., Wang, X., Mueller, D. L., Jameson, S. C., and Hogquist, K. A., 2013. Murine thymic selection quantified using a unique method to capture deleted T cells. Proc. Natl. Acad. Sci. U.S.A. **110**:4679–4684.

Su, L. F., Kidd, B. A., Han, A., Kotzin, J. J., and Davis, M. M., 2013. Virus-specific $CD4^+$ memory-phenotype T cells are abundant in unexposed adults. Immunity **38**:373–383.

Textor, J., Buytenhuijs, F., Rogers, D., Gauthier, M., Sultan, S., Wortel, I. M. N., Kalies, K., Fähnrich, A., Pagel, R., Melichar, H. J., Westermann, J., and Mandl, J. N., 2023. Machine learning analysis of the T cell receptor repertoire identifies sequence features of self-reactivity. Cell Syst. **14**:1059–1073.

Tubo, N. J., Pagan, A. J., Taylor, J. J., Nelson, R. W., Linehan, J. L., Ertelt, J. M., Huseby, E. S., Way, S. S., and Jenkins, M. K., 2013. Single naive $CD4^+$ T cells from a diverse repertoire produce different effector cell types during infection. Cell **153**:785–796.

Van Meerwijk, J. P., Marguerat, S., Lees, R. K., Germain, R. N., Fowlkes, B. J., and MacDonald, H. R., 1997. Quantitative impact of thymic clonal deletion on the T cell repertoire. J. Exp. Med. **185**:377–383.

Vrisekoop, N., Den Braber, I., De Boer, A. B., Ruiter, A. F., Ackermans, M. T., Van der Crabben, S. N., Schrijver, E. H., Spierenburg, G., Sauerwein, H. P., Hazenberg, M. D., De Boer, R. J., Miedema, F., Borghans, J. A., and Tesselaar, K., 2008. Sparse production but preferential incorporation of recently produced naive T cells in the human peripheral pool. Proc. Natl. Acad. Sci. U.S.A. **105**:6115–6120.

Yu, W., Jiang, N., Ebert, P. J., Kidd, B. A., Muller, S., Lund, P. J., Juang, J., Adachi, K., Tse, T., Birnbaum, M. E., Newell, E. W., Wilson, D. M., Grotenbreg, G. M., Valitutti, S., Quake, S. R., and Davis, M. M., 2015. Clonal Deletion Prunes but Does Not Eliminate Self-Specific alphabeta $CD8^+$ T Lymphocytes. Immunity **42**:929–941.