Predicting the optimal number of MHC proteins against positive and negative T-cell selection via clonotype-based modelling of T-cell repertoire formation

Report of a 9 month internship.
Submitted for the Masters course Bioinformatics and Biocomplexity.

Student	:	Jeroen An-Ying Saccheri
Student Number	:	0538824
Supervisor	:	Rob J. De Boer
Second Reviewer	:	Can Kesmir
Date	:	December 23, 2021

Layman's Summary

In the following article, we investigate why the number of different types of MHC molecule is so small within an individual. MHC molecules are cell surface proteins essential for the adaptive immune system, with the primary function of binding antigens and presenting them to T cells. They are known for their extreme diversity at the population level, with the genes that encode them being the most polymorphic found in vertebrates thus far. However, the number of different MHC molecules within an individual is limited: humans, for example, express 6 MHC class I molecules and between 6-8 class II molecules. MHC molecules are diverse at the population level to maximise the probability that for any pathogen, there exists a heterozygous individual within that population that can recognise it. Why the number of types present in an individual is limited is still under debate. While maturing in the thymus, T cells rearrange their T cell receptor (TCR) before undergoing positive and negative selection. The number of MHC types increases the survival rate for positive selection, but decreases the survival rate for negative selection, indicating that there is a number of MHC that maximises overall survival. Here we explore the idea that MHC diversity is limited by a need to maximise T cell survival during their development in the thymus.

The survival of a T cell during thymic selection is determined by whether it creates a functional TCR through rearrangement. We adapt a published dynamical model describing experimental data taken from homozygous mice to include the feature of considering a cell's fate to be set after TCR rearrangement, which we achieve by describing the dynamics of cells with the same fate. Using this new model, we obtain estimates of the fraction of T cells which survive positive and negative selection.

We then update a model linking the probability for T cells to survive selection to the number of different MHC molecules present in the individual, by including part of the selection process it neglected to consider. Substituting our estimates for positive and negative selection into the updated equation, we generate an estimate for the number of MHC needed to maximise survival, which resembles the true value seen in heterozygous mice. To test the robustness of our result, we built two additional adjustments to the model to more accurately mimic behaviour observed in vivo and observe if our result still holds. Under either of these changes, the optimal number of MHC types for survival remains the same. While not a proof, our results are consistent with the hypothesis that the number of MHC types is limited to optimise T cell survival during thymic selection. Greater understanding of the relationship between T cell survival and number of MHC types could provide valuable insight why an individual's MHC diversity can differ significantly between species of vertebrates.

Abstract

We adapt a previous mathematical model of T cell development with the goal of describing the dynamics of cells that have same fate during thymic selection. The original model considers the survival of a random T cell, using averages of experimental data taken from homozygous mice describing T cell counts at various stages of thymic selection. In the new 'clonotype' model we split the original model into cells surviving selection, those not surviving negative selection, and those not surviving positive selection. Solving the fractions of cells in each category from the steady state of the original model, we obtain estimates for the fraction of clonotypes surviving at each stage. These new estimates are input into another updated model that links the number of MHC molecule types to T cell survival. From this, we predict a lower and upper bound for the optimal number of MHC types to maximise survival. The true number of MHC types observed for heterozygous mice in vivo (≈ 12) falls comfortably in this estimated range ($6.5 < M < 15$). This suggests that the total number of types of MHC molecule present in an individual is influenced by a selection pressure to maximise survival of T cells during positive and negative selection. More generally, the methods used also represent a novel framework by which ODE models of populations may be split into a system of equations for distinct groups with separate outcomes.

1 Introduction

The MHC is characterised by its extreme diversity within a population. Its encoding is polygenic and codominant, and MHC genes are the most polymorphic known in vertebrates, with many loci having several hundreds -or even thousands- of alleles [3]. The currently accepted explanation for this diversity is that it maximises potential recognition of pathogens in heterozygous hosts, and in hosts expressing rare alleles [1]. However, MHC complexes are not as diverse within an individual: heterozygous humans, for example, express 6 MHC class I molecules, and between 6-8 class II molecules [3]. There is no current consensus on why diversity of MHC within an individual is limited, but one theory suggests that limiting the number of MHC types improves the proportion of T cells surviving thymic selection. By modelling the development of single positive T cells in the thymus, we aimed to investigate the impact of MHC diversity on T cell survival, and ultimately estimate the diversity of MHC types which maximises this survival probability.

Immature T cells undergo a complex process of selection in the thymus; there is extensive literature on the topic but we here provide a brief overview of the consensus understanding. Upon entering the thymus, $CD4^-CD8^-$ T cells (double negative, DN) must successfully rearrange their T cell receptor (TCR). A TCR is composed of two linking chains, α and β, which rearrange separately beginning with TCRβ. After TCRβ rearranges, it is paired with a non-rearranged preTα chain to form a pre-TCR; T cells which fail to form a pre-TCR are killed by apoptosis (β-selection). Those that succeed then attempt to rearrange the TCRα chain to form a complete TCR, and become $CD4^+CD8^+$ (double positive, DN). CD8 and CD4 are the coreceptors for TCRs to bond to either MHC Class I or Class II, respectively, thus giving double positive cells the opportunity to bind to either MHC class. However, it is likely that their TCR will be non-functional, as many rearrangements are not capable of appropriately bonding to MHC. Therefore, during positive and negative selection, the binding affinities between TCR and MHC complexes are assessed. In the thymic cortex, T cells are presented with epithelial cells expressing Class I or II MHC along with self-peptides; T cells capable of binding with suitable affinity will receive a survival signal and differentiate into $CD4^-CD8^+$ or $CD4^+CD8^-$ cells, while those which fail to bond die of neglect (positive selection) [14]. The single positive survivors migrate to the thymic medulla, where they are again presented with epithelial cells expressing MHC carrying self-antigens. Importantly, medullary epithelial cells are capable of expressing genes otherwise only seen in specific, peripheral tissues, meaning that T cells are exposed to a sample of self-antigens from around the body [7]. T cells (in either the cortex or medulla) which react with a binding affinity above a certain threshold are killed through induced apoptosis (negative selection). Finally, the surviving T cells are allowed to join the functional repertoire.

Positive selection chooses T cells that can successfully bond to the MHC of their host, which is imperative for their function, while negative selection removes those that may potentially cause autoimmunity. A higher diversity in MHC types improves the likelihood of TCR to bind to MHC, improving the probability of surviving positive selection, but also decreasing the probability of surviving negative selection (since TCR more commonly bind with an affinity above the tolerable threshold). This implies the possibility of an optimal number of MHC types such that the overall survival rate of the combined processes of positive and negative selection is maximised. If this hypothetical optimum resembles the true number of MHC types present in an individual organism, that would give an indication that an individual's MHC diversity is influenced by a need to maximise the survival probability of T cells maturing.

2 Approach

2.1 Estimating Selection

To estimate the optimal number of MHC for T cells to survive selection, we first need accurate estimates for positive and negative selection. Experimental data generally estimates the overall survival rate of thymic education to be between 2-5% [14], but accurate estimates for the independent survival rates of positive or negative selection have been harder to find. A 2014 article by Sawicka et al. [11] built a dynamic model describing T-cell maturation in mice by parameter fitting a proposed three-dimensional ODE with experimental data in mice [12]. They assumed that DN cells become pre-selection DP cells at a constant flux (cells/day), and that proliferation was sufficiently low before the SP stage as to be excluded. The three dimensions of their model are then the number of T cells in pre-DP (n_1), post-DP (n_2) and SP (n_3) stages, leading to the following system of ODEs:

$$\begin{aligned} \frac{\mathrm{d}n_1}{\mathrm{d}t} &= \phi - \phi_1 n_1 - \mu_1 n_1 \ , & \frac{\mathrm{d}n_2}{\mathrm{d}t} &= \phi_1 n_1 - \phi_2 n_2 - \mu_2 n_2 \ , \\ \frac{\mathrm{d}n_3}{\mathrm{d}t} &= \phi_2 n_2 + \lambda_3 n_3 - \xi_3 n_3 - \mu_3 n_3 \ . \end{aligned} \tag{1}$$

Parameters are given in Table 1, along with estimates for the probability of a random cell to survive positive (s_1) and negative (s_2, s_3) selection. These estimates are calculated using the probability of death in compartment i, p_i, as the rate of death μ_i divided by the sum of rates of all actions a cell could take in that compartment. The probability of surviving is then $1 - p_i$, giving $s_1 = \frac{\phi_1}{\phi_1+\mu_1}$, $s_2 = \frac{\phi_1}{\phi_1+\mu_1}$ and $s_3 = \frac{\xi_3+\lambda_3}{\xi_3+\mu_3+\lambda_3}$.

Table 1: Fitted parameters and survival probabilities found by Sawicka et al.

Parameter	Value	Description	Unit
ϕ	35.350×10^6	initial flux of DN cells into n_1	(cells/day)
ϕ_1	0.137	differentiation of cells from n_1 into n_2	(day^{-1})
ϕ_2	0.124	differentiation of cells from n_2 into n_3	(day^{-1})
ξ_3	0.151	n_3 cells exiting the thymus	(day^{-1})
μ_1	0.263	n_1 cell death by neglect	(day^{-1})
μ_2	1.370	n_2 cell death by strong TCR signal	(day^{-1})
μ_3	0.099	n_3 cell death by strong TCR signal	(day^{-1})
λ_3	0.183	cell division in n_3	(day^{-1})
s_1	34.2	n_1 survival probability	%
s_2	8.3	n_2 survival probability	%
s_3	77.1	n_3 survival probability	%

In practise, the TCR rearrangement splits the T cell repertoire into genetically distinct 'clonotypes' that can be classified into 3 or 4 fates: clones dying during positive selection, during negative selection(with two steps), and those surviving selection, so survival is largely predetermined by which clonotype a T cell belongs to. Since a given MHC will react to cells in the same clonotype in the same way, we thought it important when investigating MHC diversity's impact to mimic this behaviour, and thus adapted Equation 1 into a 'clonotype-based' model. This was done by separating the system of differential equations with multiple potential outcomes into a set of smaller separate systems, each with a unique, distinct outcome (e.g. all cells will survive until double-positive before dying). Every potential outcome of the original system is represented by its own, independent system of ODEs, thus simulating the predetermined fates of the different clonotypes. The sum of states of the new equations must be equal to the previous model, as must all fluxes in and out of each compartment. The probability of survival is then the percentage of the initial repertoire that enters the system of cells which is guaranteed to survive.

A method of building this system is demonstrated below, using a simpler two dimensional example. Consider a single population that passes through a single stage of selection (with no replication):

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = \phi - \phi_1 n_1 - \mu_1 n_1 \qquad \text{and} \qquad \frac{\mathrm{d}n_2}{\mathrm{d}t} = \phi_1 n_1 - \mu_2 n_2 \ , \tag{2}$$

where on a daily basis ϕ cells enter the pathway. This has a steady state $(\bar{n}_1, \bar{n}_2)$ at $\bar{n}_1 = \frac{\phi}{\phi_1+\mu_1}$ and $\bar{n}_2 = \frac{\phi_1 n_1}{\mu_2}$. Cells in n_1 can either pass on to n_2 at rate ϕ_1, or die at rate μ_1. Thus, the fraction of n_1 cells not dying at step 1 is $s_1 = \frac{\phi_1}{\phi_1+\mu_1}$, the survival probability.

Our aim is to define two cell types, those that always survive, and those that always die. We then want to find the relative proportion of these two cell types. A trick is to first rewrite the model with an explicit non-dimensional survival parameter, and define a differentiation rate r_1, by defining $\phi_1 = s_1 r_1$, i.e. $r_1 = \phi_1 + \mu_1$. This can be thought of as splitting the survival rate ϕ_1 into the probability a cell is chosen to survive, multiplied by the rate it takes to make that decision. Applying this trick, we see:

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = \phi - s_1 r_1 n_1 - (1 - s_1) r_1 n_1 \qquad \text{and} \qquad \frac{\mathrm{d}n_2}{\mathrm{d}t} = s_1 r_1 n_1 - \mu_2 n_2 \ ; \tag{3}$$

with the same number of parameters, only s_1 and r_1 instead of ϕ_1 and μ_1. Now we can split the model into survivors and non-survivors, N and M, by artificially setting $s_1 = 1$ or 0, for N and M respectively:

$$\begin{aligned} \frac{\mathrm{d}N_1}{\mathrm{d}t} &= f\phi - r_1 N_1 \ , \qquad \frac{\mathrm{d}N_2}{\mathrm{d}t} = r_1 N_1 - \mu_2 N_2 \\ \frac{\mathrm{d}M_1}{\mathrm{d}t} &= (1 - f)\phi - r_1 M_1 \ ; \end{aligned} \tag{4}$$

where the s_1 parameter is replaced by the fraction, f, of cells that survive the n_1 stage.

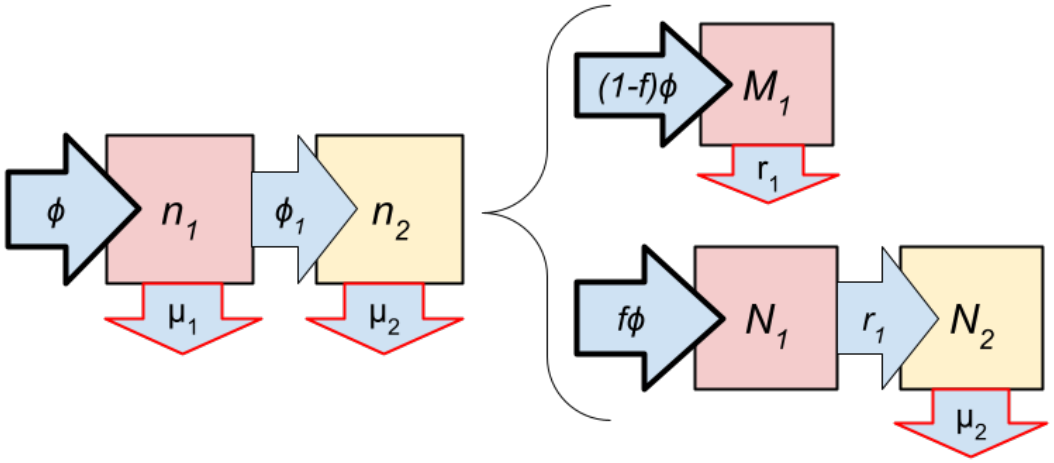


Figure 1: A diagram showing the split of a system of equations with two potential outcomes (LHS, Equation 2) into a set of systems with one potential outcome each (RHS, Equation 4). By requiring the same steady state of both models, i.e., the same total cell number in each subpopulation, and the same total flux between compartments (for example $\mu_1 \bar{n}_1 = r_1 \bar{M}_1$), we can solve the fraction, f, of initial cells fated to survive.

In fact, if $f = s_1$, this model has the same steady state as the one above, i.e. $(\bar{N}_1 + \bar{M}_1, \bar{N}_2) = (\bar{n}_1,\ \bar{n}_2)$ (see Appendix 5.1). Cells in N_1 do not die; since they move to N_2 with rate $r_1 = \phi_1 + \mu_1$, this can be thought of as 'all cells that would have died instead become N_2', while in M_1 (where death rate is also r_1) 'all cells that would have survived instead die'.

We have thus achieved our goal of splitting an original system where cells have two possible outcomes (to die in n_1 or n_2) into a set of two systems with predetermined, discrete outcomes (see Figure 1). We now look to apply to a model that includes replication. Once again consider a simple two dimensional ODE, but this time including cell division in n_1:

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = \phi + \lambda_1 n_1 - \phi_1 n_1 - \mu_1 n_1 \qquad \text{and} \qquad \frac{\mathrm{d}n_2}{\mathrm{d}t} = \phi_1 n_1 - \mu_2 n_2\ . \tag{5}$$

Considering propagation with rate λ_1 as a third possible cell action in n_1 (alongside survival ϕ_1 and death μ_1), the fraction of cells dying over time is $\frac{\mu_1}{\phi_1+\lambda_1+\mu_1}$ and hence the overall survival in this random cell model (according to the definition of Sawicka et al.[11]) is $s_1 = \frac{\phi_1+\lambda_1}{\phi_1+\lambda_1+\mu_1}$. Note that here the division rate contributes to the survival probability, i.e., s_1 increases with both ϕ_1 and λ_1. While this propagation may matter to survival for a random cell, it does not when taking the perspective of the clonotype as a whole, since all offspring should have the potential to die (with exactly the same rate as parents). In addition, since there is no limit to how many times a cell may reproduce in n_1 before leaving, the path it will take is ambiguous; this makes it difficult to separate the system into a set of systems with discrete outcomes that encompass all possibly outcomes of the original model.

Thus when constructing a clonotype model with reproduction, it was first necessary to separate the processes of selection (leaving/dying) and reproduction. We do this by considering dying/leaving as the only final outcomes for a clonotype, with reproduction then considered similarly to the flux into a compartment. With this, we can once again split into survivors and non-survivors, N and M.

$$\begin{aligned} \frac{\mathrm{d}N_1}{\mathrm{d}t} &= f\phi - r_1 N_1 + \lambda_1 N_1\ , \qquad \frac{\mathrm{d}N_2}{\mathrm{d}t} = r_1 N_1 - \mu_2 N_2 \\ \frac{\mathrm{d}M_1}{\mathrm{d}t} &= (1-f)\phi - \gamma_1 M_1\ . \end{aligned} \tag{6}$$

Note that because of the steady state of N_1, $r_1 > \lambda_1$. Since we know M_1 are guaranteed to die at step 1, we simplify the overall change as a net value: $\gamma_1 = r_1 - \lambda_1$ (where λ is necessary to match steady states to Sawicka's model). Setting $f = \frac{\phi_1}{\phi_1+\mu_1}$, as before, gives the same steady states (Appendix 5.1). However, note that f is now different from the survival probability given in Sawicka paper, as it no longer includes λ.

3 Results

3.1 An Undifferentiated Model

3.1.1 A Clonotype Model in 3 Dimensions

Using the formation of a 'clonotype-based' model, we built a set of systems of ODEs based on Equation 1. Since it has four potential outcomes, four systems of ODEs are necessary: survivors N, and non-survivors K, L and M that die at steps 1, 2 and 3 respectively. In addition, three f_i and r_i are required, one for each selection step (Appendix 5.2). As before, $f_i = \frac{\phi_i}{\phi_i+\mu_i}$ and $r_i = \phi_i + \mu_i$ (for n_3, ϕ_3 was called ξ_3 to emphasise that those cells are leaving the thymus). We again use the simplification $\gamma_3 = r_3 - \lambda_3$, to highlight that M_3 is a compartment where death is guaranteed.

$$
\begin{aligned}
\frac{\mathrm{d}N_1}{\mathrm{d}t} &= f_1 f_2 f_3 \phi - r_1 N_1 \ , & \frac{\mathrm{d}N_2}{\mathrm{d}t} &= r_1 N_1 - r_2 N_2, \\
\frac{\mathrm{d}N_3}{\mathrm{d}t} &= r_2 N_2 - r_3 N_3 + \lambda_3 N_3 \ ; & & \\
\frac{\mathrm{d}M_1}{\mathrm{d}t} &= f_1 f_2 (1-f_3) \phi - r_1 M_1 \ , & \frac{\mathrm{d}M_2}{\mathrm{d}t} &= r_1 M_1 - r_2 M_2, \\
\frac{\mathrm{d}M_3}{\mathrm{d}t} &= r_2 M_2 - \gamma_3 M_3 \ ; & & \\
\frac{\mathrm{d}L_1}{\mathrm{d}t} &= f_1 (1-f_2) \phi - r_1 L_1 \ , & \frac{\mathrm{d}L_2}{\mathrm{d}t} &= r_1 L_1 - r_2 L_2 \ ; \\
\frac{\mathrm{d}K_1}{\mathrm{d}t} &= (1-f_1) \phi - r_1 K_1 \ . & & \qquad (7)
\end{aligned}
$$

As before, we require the same steady state as the Sawicka model, as well as the same output, but it differs in reported survival probability. Since it is split into groups based on outcome instead of a random cell, overall survival probability is given based on what proportion of the initial repertoire will inevitably survive ($f_1 f_2 f_3$). As there is no propagation in n_1 or n_2, survival for the first two steps doesn't change (i.e. $f_1 = 0.342 = s_1$ and $f_2 = 0.083 = s_2$). However, we see a difference for the last value: $f_3 = 0.604 \neq s_3$.

3.1.2 Estimating Impact of MHC

With these f_i, we have the best possible estimates for overall positive and negative selection. However, to estimate the impact of the number of MHC on positive and negative selection, we require the probability of being positively or negatively selected *for one MHC*, such that we can then change the number of MHC types and see how overall survival changes. To do this, we use an update of the model

by Borghans et al.[2], extending the model by preceding it with the probability of rearranging a functional α-chain (Equation 8). β-selection will not be included, as the experimental data for our models begins at the late double negative stage, where TCRβ has already been rearranged. T cells have a one in three chance to successfully rearrange an α-chain, and if the first attempt is nonfunctional they gain a second opportunity to re-arrange on the other chromosome. Therefore, the probability of a functional α-chain rearrangement overall is $p_\alpha = \frac{1}{3} + (1 - \frac{1}{3}) \times \frac{1}{3} \approx 0.55$ [6]. Since experimental data begins at the late double negative stage and there is a time period between rearranging TCRα and becoming double posive, whether or not TCRα rearrangements have taken place is also uncertain. The probability $p_\alpha \approx 0.55$ thus represents a lower bound (where all cells had yet to rearranged TCRα), with an upper bound (all cells have already rearranged TCRα) at $p_\alpha = 1$.

With this addition, the model continues as described by Borghans et al. Assume p and n for the probability a T cell is positively or negatively selected for a single MHC, respectively. A T-cell must successfully undergo positive selection for at least 1 MHC, but avoid negative selection for all MHC, meaning the number of clones in functional repertoire R from initial repertoire R_0 can be expressed as

$$\rho = \frac{R}{R_0} = p_\alpha((1-n)^M - (1-p)^M) \ . \tag{8}$$

The total fraction of cells surviving pre-DP (n_1) is $f_1 = \alpha = 0.342$. Note this is not equivalent to p, the chance of being positively selected, as α includes the chance of surviving TCRα rearrangement, and hence $\alpha = p_\alpha(1 - (1-p)^M)$. We write the fraction of cells dying of negative selection as $\beta = 1 - f_2 f_3 = 0.950$. The overall survival is thus $\rho = \alpha(1-\beta) = R/R_0$. Using these observations to solve Equation 8 leads to:

$$p = 1 - \sqrt[M]{1 - \frac{\rho}{p_\alpha(1-\beta)}} \quad \text{and} \quad n = 1 - \sqrt[M]{1 - \frac{\rho\beta}{p_\alpha(1-\beta)}} \tag{9}$$

As we know the number of MHC in inbred mice to be $M = 6$, we solve to find find $p = 0.147$ and $n = 0.137$. The difference between these estimates is surprisingly small, ≈ 0.01. If this is accurate, it indicates the threshold within which a cell is positively selected but not negatively selected is very narrow. With these estimates of the true values, we can now re-examine Equation 8 as a function of M (Figure 2).

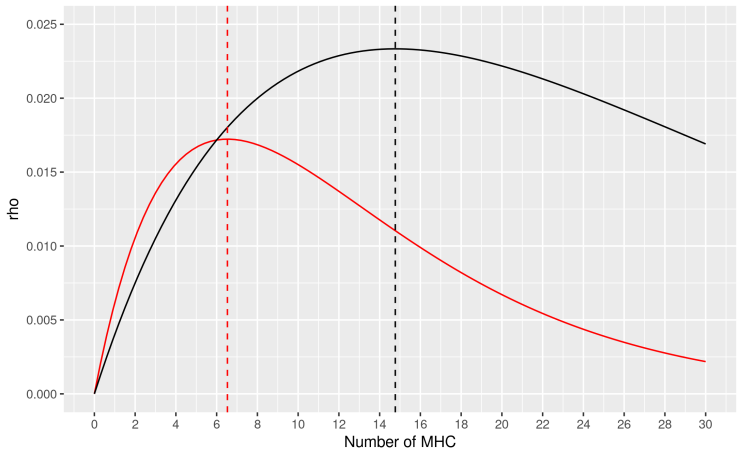


Figure 2: A graph of T cell survival 'rho' as a function of MHC diversity 'M' (Equation 8), for $p_\alpha = 0.55$ (red) and $p_\alpha = 1$ (black) showing an upper and lower bound for the optimum value of M at $6.5 < M < 15$. The true value for heterozygous mice is 12.

3.2 Extensions

3.2.1 Differentiating Between CD4$^+$ and CD8$^+$

Sawicka et al. also construct a second, more complex model that includes differentiation into CD4$^+$ or CD8$^+$ SP cells [11]. Cells in the second compartment n_2 have a chance to either become CD4$^+$ with rate ϕ_4 or CD8$^+$ with rate ϕ_8, splitting the previous undifferentiated n_3 compartment into n_4 and n_8:

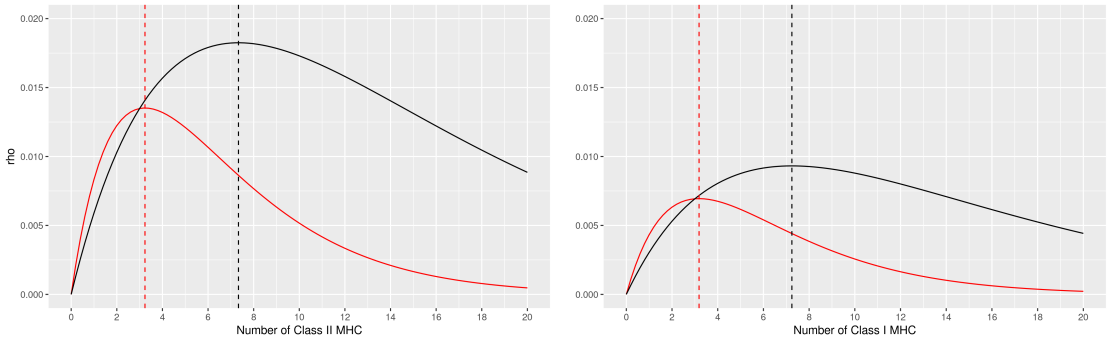
$$\begin{aligned} \frac{\mathrm{d}n_1}{\mathrm{d}t} &= \phi - \phi_1 n_1 - \mu_1 n_1 \ , & \frac{\mathrm{d}n_2}{\mathrm{d}t} &= \phi_1 n_1 - (\phi_4 + \phi_8) n_2 - \mu_2 n_2, \\ \frac{\mathrm{d}n_4}{\mathrm{d}t} &= \phi_4 n_2 + \lambda_4 n_4 - \xi_4 n_4 - \mu_4 n_4 \ , & \frac{\mathrm{d}n_8}{\mathrm{d}t} &= \phi_8 n_2 + \lambda_8 n_8 - \xi_8 n_8 - \mu_8 n_8 \end{aligned} \quad (10)$$

As before, they used linear regression and fit parameters to the same set of experimental data of T cell counts taken from homozygous mice. The rate of cells escaping n_2 should be the same between models, $\phi_4 + \phi_8 = \phi_2$. The model now has six potential outcomes: surviving as CD4$^+$ (N^4), surviving as CD8$^+$ (N^8), dying as CD4$^+$ (M^4), dying as CD8$^+$ (M^8), and dying post or pre-DP (L and K, respectively). Otherwise, the model looks much the same as Equation 7. We here show only the two groups of survivors, CD4$^+$ (N^4) and CD8$^+$ (N^8), for simplicity:

$$\begin{aligned} \frac{\mathrm{d}N_1^4}{\mathrm{d}t} &= f_1 f_2^4 f_3^4 \phi - r_1 N_1^4 \ , & \frac{\mathrm{d}N_2^4}{\mathrm{d}t} &= r_1 N_1^4 - r_2 N_2^4, \\ \frac{\mathrm{d}N_3^4}{\mathrm{d}t} &= r_2 N_2^4 - r_4 N_3^4 + \lambda_4 N_3^4 \ ; & & \\ \frac{\mathrm{d}N_1^8}{\mathrm{d}t} &= f_1 f_2^8 f_3^8 \phi - r_1 N_1^8 \ , & \frac{\mathrm{d}N_2^8}{\mathrm{d}t} &= r_1 N_1^8 - r_2 N_2^8, \\ \frac{\mathrm{d}N_3^8}{\mathrm{d}t} &= r_2 N_2^8 - r_8 N_3^8 + \lambda_8 N_3^8 \ . & & \end{aligned} \quad (11)$$

Like before, the f parameters represent the fraction of cells that survive at each step. As differentiation occurs in the second compartment n_2, there are now two paths to survival from step 2 onwards: for example f_2^4 represent the fraction of cells in n_2 that move to n_4 and f_3^4 represents the fraction that leaves n_4 to join the functional repertoire as CD4$^+$. Note that $r_2 = \phi_4 + \phi_8 + \mu_2$ is identical to before in Equation 7, since $\phi_2 = \phi_4 + \phi_8$. and thus $f_2^4 + f_2^8 = f_2$. L and K are unchanged.

Using this model, we again look to investigate the optimal MHC diversity. Since CD4$^+$ and CD8$^+$ are only affected by MHC Class II or MHC Class I, respectively, the analysis of section 3.1.2 can be performed for CD4$^+$ and CD8$^+$ independently. Estimating α_4 and β_4 for this model from f_1, f_2^4 and f_3^4 give positive and negative selection values of $p_4 = 0.273$ and $n_4 = 0.258$; doing the same for CD8$^+$ gives $p_8 = 0.273$ and $n_8 = 0.265$. As before, the difference between positive and negative selection is small, with a difference of ≈ 0.01 for both, mirroring the previous result. This narrow threshold was again surprising, and will require further investigation. Using these values for positive and negative selection in Equation 8 allows us to once again graph the effect of number of MHC on survival, this time for CD4$^+$ and CD8$^+$ independently (Figure 3).



(a) CD4$^+$ **(b)** CD8$^+$

Figure 3: Function graphs of the effect of number of MHC on the proportion of initial repertoire surviving for Class I and Class II MHC, independently. $p_\alpha = 0.55$ (red) and $p_\alpha = 1$ (black) represent a lower and upper bound considering α-rearrangement.

We see that the optimal number of MHC peaks at a value of approximately $3.25 < M < 7.25$ for both CD4 and CD8, a range which again includes the $M = 6$ for Class I and II observed in vivo in heterozygous mice (and summing to our result in section 3.1.2).

3.2.2 Exiting After Propagation

In our models, we used the same construction for propagation in the last compartment as was considered in the original Sawicka model. However, it is known that rather than having a random chance to propagate or leave at any point, T cells

leaving the thymus must go through a set number of propagations first, before they are allowed to leave[10]. We wanted to investigate whether we could add a simple version of this behaviour to the model, in the form of a required propagation before being allowed to escape. This corresponds to:

$$\frac{\mathrm{d}N_3}{\mathrm{d}t} = r_2N_2 - r_3N_3 + \lambda_3N_3 = \begin{cases} \dfrac{\mathrm{d}N_{30}}{\mathrm{d}t} = r_2N_2 - \lambda_3N_{30} \\ \dfrac{\mathrm{d}N_{31}}{\mathrm{d}t} = 2\lambda_3N_{30} - r_{31}N_{31} + \lambda_3N_{31} \end{cases} \tag{12}$$

Here, the left hand side represents the previous model, and the right hand side represents the new model with 2 generations: N_{3i}, $i = 0, 1$. The cascade has a rate of propagation λ_3, which is assumed to be the same for each generation. Only after generation 2 (the first generation of offspring, N_{31}) are T cells allowed to leave the thymus, with rate r_{31}. Naturally, this rate should be higher than the previous leaving rate r_3, since clearly $N_{31} < N_3$. Solving to match steady states to Sawicka's original model gives $r_{31} = 0.394$ (method shown in Appendix 5.3). As Sawicka et al.'s model was built with continuous propagation, propagation in the N_{31} compartment is necessary for the two models to match. In addition, it was desired to have more than 2 generations, since the expected number is 4 [10], but due to the form of Sawicka et al's model, matching steady states to a cascade with more than two generations proved mathematically impossible (5.3). Nevertheless, equation 12 still represents a functioning model with single positive T cells forced to reproduce before leaving, and since it can only be constructed without changing the values of f_i, this results in the same estimates for positive and negative selection, and the same result as in section 3.1.2.

4 Discussion

We adapted a mathematical compartment model of T cell development[11] into a clonotype-based system that allowed us to separate the initial flux of DN cells into separate groups based on outcome. By defining the survival probability as the proportion of this initial flux that went into the group guaranteed to survive, we built estimates for surviving positive and negative selection from the perspective of a clonotype. Upon these estimates, we updated a previous model [2] investigating the impact of MHC diversity on T cell survival in the thymus, which indicated that the optimal number of MHC types was $6.5 < M < 15$, resembling the true value observed in heterozygous mice. While not a proof, this gives some suggestion that MHC diversity within an individual could be limited by a need to maximise T cell survival during thymic selection, as hypothesised. In addition, we considered extensions of the model in the form of T cell differentiation and a reproductive

cascade, to see if incorporating these potentially more realistic models would affect our results. When incorporating differentiation into $CD4^+$ or $CD8^+$ into the model, the result still holds, as well as when only allowing T cells to escape the thymus after at least one division.

In the data used to parametrise Sawicka et al.'s models, T cell totals differed significantly between individual mice (as can be reasonably expected), and hence linear regression fitting procedures were used instead of parametrising directly from means of the T cell counts. In practise, this has resulted in the differentiated and undifferentiated models of Sawicka et al. having different outputs, but these are still well within their calculated standard deviations, and thus could both reasonably represent the counts observed in an individual. As such, conclusions drawn from the two models are still relevant. Because of the structure of Sawicka's original model, converting to a cascade in section 3.2.2 required some sweeping simplifications which likely do not accurately represent the true behaviour: T cells are allowed to leave after only one generation of division, it was necessary to consider division in compartment N_{31} and the rate of division is still a constant λ_3. With more experimental data of compartment times and the generational behaviour expected, a new model could be built with more generations of cell division before being allowed to leave, as well as potentially removing cell division for the final generation at the end of the cascade.

Previous attempts to investigate whether the number of MHC types optimises survival probability have come to opposing conclusions. Nowak et al.[8] found the optimum to lie close to the observed value, but an oversight in their model allowed cells to be negatively selected on MHC they were not positively selected on. Correcting for this mistake, Borghans et al.[2] then found the optimal MHC to be much higher than that observed in vivo, but their model did not consider the selection due to TCRα rearrangements (which were also neglected by Nowak). By preceding Borghans' model with the probability of a functional TCRα rearrangement, $p_\alpha = 0.55$[6], this article represents a continuation and update of the line of investigation in modelling T cell selection to investigate optimal MHC. Major differences between our work and Borghans' estimate is due to more accurate experimental data describing positive and negative selection.

The results of our research were built on experimental data from homozygous mice (M=6), but can project T cell survival rates for heterozygous mice (M=12). Experimental data of the true survival rates of T cells during thymic selection in heterozygous mice could be used to verify these projections and test the robustness of our result. Further lines of investigation could also be to continue including more realistic behaviour into the model to see if the result still holds. One such extension would be to consider T cells with two rearranged α-chains in their TCR. Sometimes the α-chain on the second chromosome of a T cell rearranges before

the functionality of the first rearrangement has been verified, resulting in a T cell with two rearranged TCRα. Research suggests as many as 25% of T cells in the functional repertoire in our repertoire possess this trait[5]. The presence of this extra TCRα slightly increases the probability of binding to a MHC, and thus could affect the observed optimum. Since ordinary differential equations on this topic tend to resemble the Sawicka model[10], the methods of splitting a system of equations into a set of smaller systems described in this article could likely be adapted to investigate other models of T cell development.

Code

All analyses were performed in R[9], using the tidyverse package[13] and Grind[4], an R-script built for phase plane analysis. Code used for analyses can be found at `https://github.com/Jeroenioeni/MHC` in R markdown format.

References

[1] José A. M. Borghans, Can Keşmir, and Rob J. De Boer. MHC diversity in Individuals and Populations. In Darren Flower and Jon Timmis, editors, *In Silico Immunology*, pages 177–195. Springer US, Boston, MA, 2007.

[2] José A. M. Borghans, André J. Noest, and Rob J. De Boer. Thymic selection does not limit the individual MHC diversity. *European Journal of Immunology*, 33(12):3353–3358, December 2003.

[3] Jr Charles A Janeway, Paul Travers, Mark Walport, and Mark J. Shlomchik. The major histocompatibility complex and its functions. *Immunobiology: The Immune System in Health and Disease. 5th edition*, 2001.

[4] Rob J. de Boer. Grind. `https://tbb.bio.uu.nl/rdb/grindR.html`. Accessed: 2021-11-30.

[5] Peter C de Greef, Theres Oakes, Bram Gerritsen, Mazlina Ismail, James M Heather, Rutger Hermsen, Benjamin Chain, and Rob J de Boer. The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *eLife*, 9:e49900, March 2020.

[6] Andreas Krueger, Natalia Ziętara, and Marcin Łyszkiewicz. T Cell Development by the Numbers. *Trends in Immunology*, 38(2):128–139, February 2017.

[7] Amy E Moran and Kristin A Hogquist. T-cell receptor affinity in thymic development. *Immunology*, 135(4):261–267, April 2012.

[8] M. A. Nowak, K. Tarczy-Hornoch, and J. M. Austyn. The optimal number of major histocompatibility complex molecules in an individual. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10896–10899, November 1992.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[10] Philippe A. Robert, Heike Kunze-Schumacher, Victor Greiff, and Andreas Krueger. Modeling the Dynamics of T-Cell Development in the Thymus. *Entropy*, 23(4):437, April 2021.

[11] Maria Sawicka, Gretta L. Stritesky, Joseph Reynolds, Niloufar Abourashchi, Grant Lythe, Carmen Molina-París, and Kristin A. Hogquist. From pre-DP, post-DP, SP4, and SP8 Thymocyte Cell Counts to a Dynamical Model of Cortical and Medullary Selection. *Frontiers in Immunology*, 5, 2014.

[12] G. L. Stritesky, Y. Xing, J. R. Erickson, L. A. Kalekar, X. Wang, D. L. Mueller, S. C. Jameson, and K. A. Hogquist. Murine thymic selection quantified using a unique method to capture deleted T cells. *Proceedings of the National Academy of Sciences*, 110(12):4679–4684, March 2013.

[13] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[14] Andrew J. Yates. Theories and Quantification of Thymic Selection. *Frontiers in Immunology*, 5:13, February 2014.

5 Appendix

5.1 Matching Steady States

We look to show that for the simplest case, steady states match if $f = \frac{\phi_1}{\phi_1+\mu_1}$. Take the simple model without replication seen in Equation 2 and its separated form shown in Equation 4:

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = \phi - \phi_1 n_1 - \mu_1 n_1 \quad \text{and} \quad \frac{\mathrm{d}n_2}{\mathrm{d}t} = \phi_1 n_1 - \mu_2 n_2 \ ; \tag{2}$$

$$\frac{\mathrm{d}N_1}{\mathrm{d}t} = f\phi - r_1 N_1 \ , \qquad \frac{\mathrm{d}N_2}{\mathrm{d}t} \ = \ r_1 N_1 - \mu_2 N_2$$

$$\frac{\mathrm{d}M_1}{\mathrm{d}t} = (1-f)\phi - r_1 M_1 \ . \tag{4}$$

Equation 2 has steady state $\bar{n}_1 = \frac{\phi}{\phi_1+\mu_1}$ and $\bar{n}_2 = \frac{\phi_1 \bar{n}_1}{\mu_2} = \frac{\phi_1 \phi}{\mu_2(\phi_1+\mu_1)}$, and Equation 4 has steady state $\bar{N}_1 = \frac{f\phi}{r_1}$, $\bar{N}_2 = \frac{r_1 \bar{N}_1}{\mu_2} = \frac{f\phi}{\mu_2}$ and $\bar{M}_1 = \frac{(1-f)\phi}{r_1}$.

We now attempt to match these steady states. Since $r_1 = \phi_1 + \mu_1$, $\bar{N}_1 + \bar{M}_1 = \bar{n}_1$ for any f. $\bar{N}_2 = \bar{n}_2$ iff. $f = \frac{\phi_1}{\phi_1+\mu_1}$, which was to be demonstrated.

5.2 General Framework

We here show a general framework of a three-dimensional ODE allowed to propagate at every step with rate λ_i:

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} \ = \ \phi + \lambda_1 n_1 - \phi_1 n_1 - \mu_1 n_1, \qquad \frac{\mathrm{d}n_2}{\mathrm{d}t} = \phi_1 n_1 + \lambda_2 n_2 - \phi_2 n_2 - \mu_2 n_2,$$

$$\frac{\mathrm{d}n_3}{\mathrm{d}t} \ = \ \phi_2 n_2 + \lambda_3 n_3 - \phi_3 n_3 - \mu_3 n_3. \tag{13}$$

This has steady states $\bar{n}_1 = \frac{\phi}{\phi_1+\mu_1-\lambda_1}$, $\bar{n}_2 = \frac{\phi_1 n_1}{\phi_2+\mu_2-\lambda_2}$ and $\bar{n}_3 = \frac{\phi_2 n_2}{\phi_3+\mu_2-\lambda_3}$.
We split into K, L, M and N, the survivors of different stages, with N as the only group that survives to leave the system.

$$
\begin{aligned}
\frac{\mathrm{d}N_1}{\mathrm{d}t} &= f_1 f_2 f_3 \phi - r_1 N_1 + \lambda_1 N_1 \ , & \frac{\mathrm{d}N_2}{\mathrm{d}t} &= r_1 N_1 - r_2 N_2 + \lambda_2 N_2, \\
\frac{\mathrm{d}N_3}{\mathrm{d}t} &= r_2 N_2 - r_3 N_3 + \lambda_3 N_3 \ ; & & \\
\frac{\mathrm{d}M_1}{\mathrm{d}t} &= f_1 f_2 (1 - f_3) \phi - r_1 M_1 + \lambda_1 M_1 \ , & \frac{\mathrm{d}M_2}{\mathrm{d}t} &= r_1 M_1 - r_2 M_2 + \lambda_2 M_2, \\
\frac{\mathrm{d}M_3}{\mathrm{d}t} &= r_2 M_2 - \gamma_3 M_3 \ ; & & \\
\frac{\mathrm{d}L_1}{\mathrm{d}t} &= f_1 (1 - f_2) \phi - r_1 L_1 + \lambda_1 L_1 \ , & \frac{\mathrm{d}L_2}{\mathrm{d}t} &= r_1 L_1 - \gamma_2 M_2 \ ; \\
\frac{\mathrm{d}K_1}{\mathrm{d}t} &= (1 - f_1) \phi - \gamma_1 K_1. & & \qquad (14)
\end{aligned}
$$

For simplification, $(\phi_i + \mu_i)$ has been replaced by r_i and for compartments with guaranteed death, $r_i - \lambda_i$ has been replaced by net change, γ_i. We require several fractions f_i, one for each selection step: f_1, f_2 and f_3. Choosing $f_i = \phi_i/(\phi_i + \mu_i)$, the survival probability in the absence of replication, results in the steady states matching to the full model above. This can be trivially shown using the same technique as Appendix 5.1. Thus, in contrast to the cell-based model, the λ_i drop out from the survival probabilities.

5.3 Parametrising the Cascade

We look to find parameters for λ_{3x} and r_{33} such that for Equation 12, the left hand side is equal to the sum of the right hand side. Since totals are the same, $N_3 = \sum N_{3i}$, and so $\mathrm{d}N_3/\mathrm{d}t = \sum \mathrm{d}N_{3i}/\mathrm{d}t$:

$$
\begin{aligned}
r_2 N_2 - r_3 N_3 + \lambda_3 N_3 \quad &= \quad r_2 N_2 - \lambda_3 N_{30} + \\
&\quad\ 2\lambda_3 N_{30} - r_{31} N_{31} + \lambda_3 N_{31} \qquad (15) \\
&= \quad r_2 N_2 + \lambda_3 (N_{30} + N_{31}) - r_{31} N_{31}
\end{aligned}
$$

Since $N_3 = \sum N_{3i}$, this simplifies to $r_3 N_3 = r_{31} N_{31}$. This, in fact, is also one of the required properties from the beginning: that the outputs should match. Using the steady states $\bar{N}_3 = \frac{r_2 \bar{N}_2}{r_3 - \lambda_3}$ and $\bar{N}_{31} = \frac{2\lambda_3 \bar{N}_{30}}{r_{31} - \lambda_3} = \frac{2 r_2 \bar{N}_2}{r_{31} - \lambda_3}$, we solve $r_3 \bar{N}_3 = r_{31} \bar{N}_{31}$, an equation with only one unknown, to find $r_{31} \approx 0.394$. Note that this only changes $r_3 1$, and not the survival parameters f_i.

We now investigate whether it is possible to build a cascade with more than one division step. Consider a cascade with n divisions:

$$\begin{aligned}
\frac{\mathrm{d}N_{30}}{\mathrm{d}t} &= r_2 N_2 - \lambda_3 N_{30} \\
\frac{\mathrm{d}N_{3i}}{\mathrm{d}t} &= 2\lambda_3 N_{3(i-1)} - \lambda_3 N_{3i} \qquad\qquad i \in \{1, (n-1)\} \\
\frac{\mathrm{d}N_{3n}}{\mathrm{d}t} &= 2\lambda_3 N_{3(n-1)} - r_{3n} N_{3n} + \lambda_3 N_{3n};
\end{aligned} \tag{16}$$

By the same process as Equation 15, this simplifies to $r_3 N_3 = r_{3n} N_{3n}$. The steady state of N_{3n} is $\bar{N}_{3n} = \frac{2^n r_2 N_2}{r_{3n} - \lambda_3}$. For $n > 1$, solving $r_3 \bar{N}_3 = r_{3n} \bar{N}_{3n}$ results in a negative value for r_{3n}, which is impossible. Therefore, it is not possible to build a cascade in this form from Sawicka's model for more than one division step.

5.4 Preliminary Project

We began investigating the formation of T cell functional repertoires and the parameters guiding thymic selection in an earlier project involving the 'minifish', genus *Paedocypris*. Due to the fish's tiny size, it has very few lymphocytes, and it was expected that this would result in a need for a novel solution to the problem of how to develop a sufficiently specific adaptive immune system. Our findings indicated that instead, even for this extreme example, a sufficiently specific adaptive immune system is possible to construct using the same simple toy models made to apply to humans and mice. The resulting report is attached below.

Response probabilities of the adaptive immune system in a small vertebrate

Abstract

A recent study used one of the smallest known vertebrates, *Paedocypris* sp.”Singkep”, as a model organism due to its low lymphocyte count. Their estimates of lymphocyte count and proteome size are used here to investigate the claim that smaller vertebrates should possess less reactive adaptive immune systems due to smaller lymphocyte repertoires. A probabilistic model shows instead that even for the smallest known vertebrates, the adaptive immune system is sufficiently reactive, especially towards larger proteins that present more epitopes.

Introduction

Antigen receptor repertoires are the foundation of the adaptive immune system, but their form, structure and constraints are still not fully understood. The sheer size of most vertebrate repertoires makes it impractical to sequence a large proportion of an individual’s total lymphocyte population, and high heterogeneity between lymphocyte clonotypes makes it challenging to represent an individual’s total repertoire based on a partial sample. Giorgetti et al.[1] aimed to circumvent this sampling problem by studying one of the smallest known vertebrates: a cyprinid ‘minifish’, *Paedocypris* sp. “Singkep”. They estimated the fish to possess ~37,000 T cells in a mature adult, hence making it feasible to sequence the majority of clonotypes in an individual and gain a representative sample.

However, the low lymphocyte count of *Paedocypris* raised a fundamental question about how its immune system functions. The adaptive immune system trains itself to not react to self-proteins by a process known as self-tolerance: lymphocytes held in lymphoid organs are introduced to self-proteins, and destroyed if they bind with strength above a tolerable threshold[2]. Since it is reasonable to assume an upper bound on the total number of lymphocytes an organism can produce, the number of self-proteins constrains the probability of an individual lymphocyte reacting to an epitope (also known as its specificity), as with more self-proteins there are more responses to avoid. *Paedocypris* produces approximately 12,000 self-proteins, almost as many as larger vertebrates like humans (which produce ~20,000). Under similar constraints, it can be reasoned that the specificity per lymphocyte of *Paedocypris* should also be roughly the same to avoid autoimmunity. Because the specificity per lymphocyte must be similar, but the size of the repertoire is several orders of magnitude smaller ($4x10^{11}$ in humans[3] vs $4x10^{4}$ in *Paedocypris*[1])), we hypothesize that a simple probabilistic model of *Paedocypris*’s immune system would estimate a very low probability to react to a foreign epitope, therefore requiring a more complex explanation. We used the data of Giorgetti et al. to make a toy model and investigate this claim.

Model and Results

The model below[4] describes the formation of a functional repertoire (Eq. 1) and its probability to react to a given epitope (Eq. 2). In (1), an initial repertoire of size R_0 is introduced to S self epitopes. Lymphocytes react with specificity p and self tolerance kills reacting lymphocytes to leave a functional repertoire of size R. In (2), we see the recognition probability P_i : the probability a functional repertoire of size R and specificity p recognises a given epitope i.

$$(1) \qquad R = R_0 \, (1 - p)^S \qquad\qquad (2) \qquad P_i = 1 - (1 - p)^R$$

Since the specificity is significantly smaller than 1, an exponential approximation can be used : $(1 - p)^n \simeq e^{-pn}$. Incorporating this into the formulae above and differentiating with respect to specificity, we find that the recognition probability is optimised at specificity $\hat{p}=1/S$. Note that this has been done without referencing any particular species. Across all vertebrates, this simple model indicates that the optimal specificity will be inversely proportional to the number of self epitopes (as had been intuitively expected).

For the optimum specificity, we can estimate all the parameters to model the immune system of *Paedocypris*. As *Paedocypris* has ~12,000 self-proteins, similar to humans, we assume that as in humans S is roughly 10^5, and thus the optimal p is $\hat{p}=10^{-5}$. R is ~37,000, the recorded number of T cells. Using this value of R and p results in a recognition probability per epitope of $P_i = 0.30$.

Using the same model on humans or mice gives a recognition probability per epitope of close to 1 (see Fig.1 comparing humans to *Paedocypris*). Is $P_i = 0.30$ high enough for an immune system to function? If a foreign pathogen presents 10 epitopes, a recognition probability of $P_i = 0.30$ means the chance of recognising such a pathogen is greater than 97% ($1-0.7^{10}$).

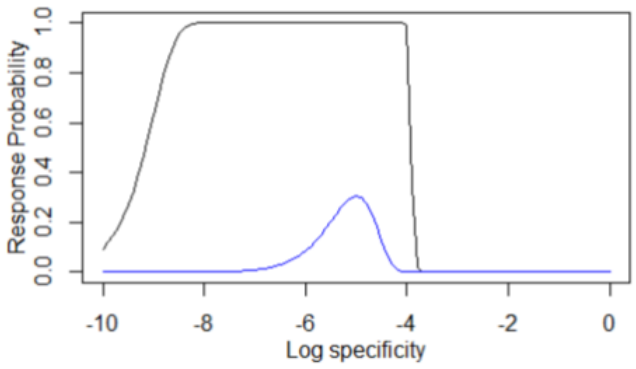


Figure 1
A plot showing the response probability (P_i) against specificity (p, in a logarithmic scale) for *H. sapiens* (black) and *Paedocypris* (blue), with a fixed initial repertoire size.. Note both have optimal regions around $p=10^{-5}$ (though the range possible for *H. sapiens* is significantly larger).

Discussion

Known pathogens present several epitopes[5]. The genome of HIV, one of the smallest known viruses, encodes for 15 viral proteins[6]. Even if each protein was just 1 epitope, HIV would present 15 epitopes as a whole. Our hypothetical pathogen with 10 epitopes thus represents a lower bound. In the simple model, any pathogen presenting over 10 epitopes will be recognised with a probability of above 97%, which is sufficient for a functional immune system. While this would drop sharply for a hypothetical pathogen that only presents a couple of epitopes, such a pathogen is currently unknown.

The result that an organism has a functional immune system may seem somewhat obvious, but it is not trivial. *Paedocypris* lies at the extreme of what is possible for vertebrates, being a very small organism with a relatively large proteome, so the result should hold across all vertebrates. If a recognition probability per epitope of 30% is sufficient, is the ~100% recognition of humans and mice something that has been positively selected for, or is it a byproduct of a system maximising against other constraints? We would argue the latter.

Having a larger body may necessitate a higher recognition probability due to the time it takes for a lymphocyte to come into contact with an epitope. Lower recognition probabilities could increase the time taken for a system coming into contact with an epitope to recognise it. This would make a more significant difference in a large organism compared to something small like *Paedocypris*. However, experiments in mice have indicated newly introduced epitopes are brought to primary lymphoid organs like the thymus relatively quickly[7]; there are mechanisms in place to ensure that a new epitope is presented to different lymphocyte clonotypes as quickly as possible[8]. Therefore, while this is something to consider, it is unlikely to be a main constraint.

It has previously been argued, with reference to humans and mice, that the structure of the adaptive immune system is more focused on avoiding autoimmunity than maximising probability to respond to epitopes[4]. This result supports that claim, as it shows that a specificity constrained by the number of self-epitopes can provide a sufficient immune response even in this extreme case. If that is true, this should hold for all vertebrates, so the adaptive immune system can always be constructed to avoid autoimmunity without needing to greatly compromise the recognition probability.

Bibliography

1. Giorgetti, O. B. *et al.* Antigen receptor repertoires of one of the smallest known vertebrates. *Sci. Adv.* **7**, (2021).
2. Alberts, B. *et al. Molecular biology of the cell*. (Garland Science, 2014). doi:10.1201/9780203833445.
3. Jenkins, M. K., Chu, H. H., McLachlan, J. B. & Moon, J. J. On the composition of the preimmune repertoire of T cells specific for Peptide-major histocompatibility complex ligands. *Annu. Rev. Immunol.* **28**, 275–294 (2010).
4. De Boer, R. J. & Perelson, A. S. How diverse should the immune system be? *Proc. Biol. Sci.* **252**, 171–175 (1993).
5. Assarsson, E. *et al.* A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection. *J. Immunol.* **178**, 7890–7901 (2007).
6. Frankel, A. D. & Young, J. A. HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67**, 1–25 (1998).
7. Miao, H. *et al.* Quantifying the early immune response and adaptive immune response kinetics in mice infected with influenza A virus. *J. Virol.* **84**, 6687–6698 (2010).
8. Yan, F. *et al.* Thymic function in the regulation of T cells, and molecular mechanisms underlying the modulation of cytokines and stress signaling (Review). *Mol. Med. Report.* **16**, 7175–7184 (2017).