Modelling peripheral homeostasis of naive CD_4^+ T-cells with a static distribution of lifetimes for thymic emigrants

Student: Martin O'Reilly†; Supervisors: Robin Callard and Iren Bains

T-cells play a crucial regulatory role in the immune system. Understanding the mechanisms underlying the homeostasis of the healthy T-cell pool is expected to provide insight into the apparent deregulation of this process in many auto-immune diseases, including HIV. Naive CD_4^+ T-cells form a significant fraction of the overall T-cell pool, and it has been observed that the average lifetime of these cells increases with the age of an individual. This case essay explores a simple potential mechanism for this increase in lifetime. This mechanism assumes a static distribution of lifetimes for all T-cells newly exported from the Thymus, corresponding to a static distribution of "fitness" for survival. The model hypothesises that the accumulation of long-lived survivors in the T-cell pool over time will result in the observed increase in average lifetime. Initial evaluation of the model determines that it does not explain the observed increase in lifetime in its current form. However, potential adjustments to the model are suggested that might result in a closer fit to the data.

Introduction

T-cells play a crucial regulatory role in the immune system, and are essential in order for the body to mount a proper immune response to a variety of pathogens. Abnormal T-cell function is implicated in a range of auto-immune diseases, including HIV. Therefore research into the development and maintenance of the T-cell pool comprises a large and active field. Much remains unknown about how T-cell numbers are suppressed by infections such as HIV, and what effect this suppression has on the function of the T-cell pool. For example, it is not clear whether the variety of the T-cell pool (and therefore the range of potential infections it can recognise) is maintained when its numbers are reduced by infection. It is hoped that gaining a better understanding of how the healthy T-cell population is generated and maintained will provide insight into how the T-cell pool is altered by, and recovers from, infection.

One interesting observation of T-cell development is that the average lifetime of the naive $CD4^+$ subset of T-cells seems to increase with the age of an individual, at least until adulthood. This case essay explores a simple potential mechanism for this increase in lifetime. This mechanism assumes a static distribution of lifetimes for all T-cells newly exported from the Thymus, corresponding to a static distribution of "fitness" for survival. The model hypothesises that the accumulation of long-lived survivors in the T-cell pool over time will result in the observed increase in average lifetime.

Biological background

The immune system comprises two semi-autonomous parts: the innate and adaptive systems.

The innate immune system

The innate system evolved first, and utilises a small repertoire of generic receptors and anti-microbial agents to recognise and attack a wide range of pathogens. The anti-microbial agents include both short sequences of amino acids and fully-fledged proteins, including the important *complement* proteins. These generic receptors and agents are specific for highly conserved groups of molecules that are associated with *broad subsets* of pathogens. Crucially, they are also genetically "hard-coded" and thus are inherited directly (with reproductive cross-over and mutation) from parent to offspring. This means that the innate system's repertoire of receptors and agents is fixed at birth and cannot change in response to exposure to a new pathogen. Changes to the innate repertoire can only occur on an evolutionary timescale. In addition the number of different receptor or agent patterns that can be genetically coded is limited, resulting in the limited repertoire of the innate system. Given this limited repertoire, it is unsurprising that receptors and agents which respond to a wide range of pathogens were selected for during the evolution of the innate system.

The adaptive immune system

The adaptive system evolved later and it utilises a much larger repertoire of much more specialised receptors and agents, which are highly specific to molecules associated with *individual* pathogens. Unlike the genetically hard-coded repertoire of the innate system, the receptors and agents of the adaptive system are randomly generated during production. This results in a much wider repertoire than could ever be genetically hard-coded. Of course a large number of the randomly generated receptors and peptides will be non-functional, or even harmful to the body (provoking an *auto-immune response*). However, quality control mechanisms during production weed these out, ensuring a high level of function across the diverse repertoire.

B-cells

In the adaptive system, the anti-microbial agents are *antibodies*. Produced in B-cells and present on B-cell membranes, antibodies are also secreted into the bloodstream during an immune response. Antibodies are immunoglobulin molecules, and a single B-cell expresses only *one* specific antibody on its membrane. This same antibody is secreted by a subset of a B-cell's offspring when it is activated during an immune response. B-cell antibodies are highly specific to molecules associated with a single *individual* pathogen. These molecules are called *antigens*. During B-cell production the antibody-encoding DNA of each cell undergoes random recombination and mutation, producing a unique antibody for each cell and resulting in a diverse antibody repertoire.

Approximately 10^7 B-cells are produced in the bone marrow each day. However, only 10% of these leave the bone marrow. The remaining 90% are selected against in the marrow and destroyed. This

† MRes student at UCL CoMPLEX (Centre for Mathematics and Physics in the Life Sciences and Experimental Biology).
Email: martin.o'reilly@ucl.ac.uk

may be because their antibodies are malformed or because they react strongly to molecules naturally produced by the body (self-antigens). A strong reaction to self-antigens could cause the immune system to mount an immune response against its own cells and is thus very undesirable. As not all self-antigens are found in the bone marrow, it is likely that some further selection occurs in the periphery (blood, organs, lymphatic system etc). Once in the periphery, newly exported B-cells live for a short time (a few weeks) unless they are activated by antigen. B-cell activation requires two signals. The first is the cross-linking of two membrane-bound antibodies by an antigen. The second signal is usually provided by a helper (T_H) T-cell binding to an antigen-presenting class II Major Histocompatibility Complex (MHC[1]) on the membrane of the B-cell. However, this signal can also be provided by a class of antigens known as *thymus-independent antigens*. Nonetheless, most B-cell activation requires T_H-cell interaction. These two activation processes are illustrated in figure 1.

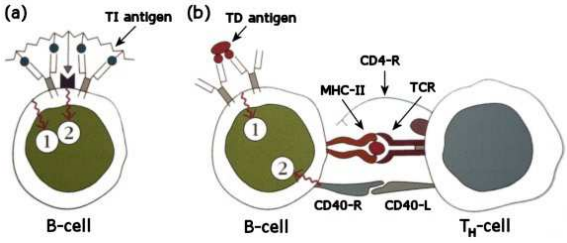


Figure 1: B-cell activation. (a) Thymus-independent activation, where a single antigen provides both activation signals. (b) Thymus-dependent activation, where both antigen and T_H-cell are required for activation. MHC-II is an antigen-presenting class II MHC; TCR is the T-cell receptor; CD4-R is the CD4 co-receptor; CD40-R and CD40-L are the CD40 receptor and ligand respectively. *[Source: Adapted from Kindt et al, 2007]*

T-cells

The receptors in the adaptive system are *T-cell receptors* (TCRs). These are membrane-bound receptors and, as for B-cell antibodies, these receptors are highly specific to molecules associated with a single *individual* pathogen. However, unlike B-cell antibodies, TCRs do not interact with free antigen. Instead, the antigen must be presented to the TCR by a Major Histocompatibility Complex (MHC) on an *antigen presenting cell*. MHCs are only expressed on the membrane of the body's own cells. therefore TCRs are specific for a combination of antigen *and* self.

There are two types of MHC, class I and class II, and the class of MHC a T-cell requires for activation distinguishes cytotoxic (T_C) T-cells from helper (T_H) T-cells. These cells also have different co-receptors which must also be engaged for T-cell activation. T_C-cells require antigens to be presented via class I MHCs and express the CD8 co-receptor, thus they are also known as $CD8^+$ T-cells. T_H-cells require antigen to be presented via class II MHCs and express the CD4 receptor, thus they are also known as $CD4^+$ T-cells.

$CD8^+$ and $CD4^+$ T-cells play two different roles in the adaptive immune system. Activated $CD8^+$ T-cells are known as cytotoxic T-lymphocytes (CTLs) and emit cell-destroying chemicals called *cytotoxins*. Almost all nucleated cells in the body express class I MHCs. Therefore a CTL can interact with and destroy any cell in the body that presents the antigen specific to its TCR, as long as the right co-stimulatory immune signals are present. For most normal body cells, these co-stimulatory signals are only present when the cell is infected or cancerous. However, in order for an inactive (or *naive*) CD8+ T-cell to become a CTL it must first be activated. Evidence is emerging that it is not sufficient for a naive CD8+ cell to interact with just any antigen presenting cell for activation. Rather activation may require interaction by so-called *licensed* antigen-presenting cells (APCs). Licensing may require interaction between the antigen-presenting cell and $CD4^+$ T-cells (Kindt et al, 2007). Only a small set of *professional* APCs express both the class I MHCs required for interaction with CD8+ T-cells and the class II MHCs required for interaction with CD4+ T-cells. These APCs, such as dendritic cells and macrophages, are generally cells of the innate immune system and this is one of several points the two systems interact. While the role of CD4+ T-cells in *licensing* is still not clear, such a roll in CD8+ T-cell activation mirrors the role of CD4+ T-cells in B-cell activation.

As has been described already, and as suggested by their name, $CD4^+$ helper T-cells play a facilitator role in the immune system. They are critical for the majority of B-cell activation and it seems they may also be important for $CD8^+$ T-cell activation. In addition to interacting with membrane proteins on other immune cells, activated $CD4^+$ T-cells also secrete a range of helper chemical called *cytokines*. These play an important role in regulating the immune response. Certain cytokines are required for B-cell activation and APC licensing, while others recruit a range of different immune cells to the site of an infection. $CD4^+$ TCRs only interact with antigens presented via class II MHCs. These are only present on other immune system cells such as dendritic cells, macrophages and B-cells. It appears that dendritic cells play the major role initial T-cell activation, while B-cells may play an important role in maintaining T-cell activity when the levels of pathogen fall.

T-cell production

T-cells are produced in the thymus (which is where they get their name) and, like B-cells in the bone marrow, undergo both positive and negative selection prior to release into the periphery. The T-cell receptor (TCR) is composed of two parts, called α and β chains[2]. T-cells in the thymus are initially double negative (DN), expressing neither CD4 nor CD8 co-receptors, and have not yet developed a TCR. The β chain of the TCR is assembled first, and its coding DNA is randomly re-arranged prior to assembly, resulting in a unique chain. Epithelial cells in the thymus express both class I and II MHCs (likely presenting a range of self-antigens) and T-cells

[1] MHCs are cell membrane complexes that can present short snippets of peptide on the surface of a cell.

[2] A small proportion (~5%) of T-cells have TCRs composed of γ and δ chains. These TCRs do not seem to require antigen to be presented via an MHC and can interact directly with peptides and proteins. These T-cells seem to be concentrated in the region of the gut and therefore may play a role in maintaining the integrity of the gut barrier to pathogen transfer. The description of T-cell production above is focussed on $\alpha\beta$ T-cells for simplicity. $\gamma\delta$ chain re-arrangement completes at the same stage as β chain re-arrangement (prior to MHC selection) and $\gamma\delta$ is then complete.

possessing a β chain which does not interact sufficiently with these will fail to survive. Following this first round of selection, the surviving T-cells undergo a period of significant division, producing considerable sub-populations of cells expressing each of the β chains that survived selection. This ensures that each successful β chain is paired with many different α chains. This both increases the TCR variation in the T-cell repertoire and increases the chance that each successful β chain is paired with at least one successful α chain and is not lost to the repertoire.

Following this proliferation stage the α chain is assembled, utilising the same DNA re-arrangement as β chain assembly to ensure unique αβ chain combinations. Once both chains of the TCR have formed, the T-cells become double positive (DP), expressing both CD4 and CD8 co-receptors. At this point there is a further round of selection. Kindt et al (2007) states that this round only selects against T-cells which interact too strongly with either MHC alone or MHC+self-antigen combinations. However, it seems likely that some of the randomly produced αβ chain combinations will be either non-functional or interact too weakly with MHC or MHC+self-antigen and therefore be selected against. Either during this selection or immediately after, T-cells differentiate into single positive (SP) $CD4^+$ or $CD8^+$ T-cells and the surviving cells are released into the periphery. As with B-cells, the majority of prospective T-cells do not make it though selection, with an estimated 95-98% dying in the thymus. However, the remaining 2-5% still represents a thymic export of between $2x10^8$ and $7x10^8$ T-cells per day for $CD4^+$ T-cells alone[3] (Bains et al, in press).

As the terms MHC+antigen and MHC+self-antigens are somewhat unwieldy, the rest of this case essay will leave the MHC element implicit and simply use antigen and self-antigen respectively when discussing T-cell interactions.

T-cell homeostasis

Once in the periphery, T-cells require a certain amount of interaction to stay alive and also undergo a certain level of division. A cell's descendants will all share the same TCR. Such a group of cells with a common TCR a known as a clonotype. There are an estimated 10^7 different clonotypes in the human periphery (Arstila et al, 1999). With $\sim10^{11}$ naive T-cells in the periphery (Bain et al data), this infers that each clonotype has ~10,000 member T-cells. This is substantially higher than the clonotype size of 20-200 reported for mice (Surh and Sprent, 2008). While, the mouse figure represents a consensus range drawn from several studies and the human figure is from a single study, it should be noted that results from mouse studies often cannot be directly applied to humans. In this case, the difference in number of cells per clonotype is likely to reflect the difference in total T-cell populations between the two species, resulting from their substantially different physical sizes. Total T-cell numbers appear to be limited by the available levels of the cytokine IL-7. This seems to be a general stimulus that doesn't favour any particular clonotype. However, T-cell survival also requires the interaction of TCR with self-antigen (Surh and Sprent, 2008). A wide variety of self-antigens are presented by various cells in the immune system, sufficient to stimulate a similarly wide variety of TCRs. This survival interaction mirrors the process of selection in the thymus. While this interaction is low affinity, there will be a limited number of clonotypes that can receive a survival stimulus from a particular self-antigen. This results in competition between clonotypes with similar TCRs. Thus the number of sustainable clonotypes is likely to be primarily a function of the level of variety in the range of presented self-antigens and the level of overlap in the sets of self-antigens with which the TCRs of the various clonotypes interact. It is likely that the variety of self-antigens and TCRs is independent of the host's physical size, as this variety is generated by random DNA recombination at a microscopic level. Thus it is not surprising that the typical clonotype population varies significantly from mouse to human.

As mentioned above, the survival of naive T-cells requires both a general IL-7 stimulus and a more TCR-specific self-antigen stimulus. The competition for these survival stimuli limits the overall naive population and, with $\sim10^8$ T-cells released from the thymus each day, a similarly large number of T-cells must therefore die each day. The proliferation of naive T-cells also requires the same IL-7 and TCR-specific survival stimuli, and it has been shown that the TCR-specific ligands controlling T-cell proliferation are similar to those controlling T-cell selection in the thymus.

The cyton model

Hawkins et al (2007) studied the characteristics of B-cell cell survival and proliferation and concluded that cell death and cell division were independently controlled within the cell. From their measurements they determined that cell time-to-die and time-to-divide were both log-normally distributed[4]. They propose that an individual cell draws a time-to-die and a time-to-divide from this distribution, and uses this two set independent "countdown timers" for death and division. The timer that runs out first determines the fate of the cell. If this fate is death, the cell dies. If this fate is division, then both clocks are reset and a new time-to-die and time-to-divide are drawn. They coined the term cyton for the abstract cellular "machinery" that controls this process.

It is difficult to determine how applicable these in-vitro B-cell findings are to the homeostasis of in-vivo T-cells. Firstly, the cells in the cyton study were being flooded with division stimuli, dividing every 9 hours after a 30-50 hour variable delay to first division. In contrast, it is estimated that in-vivo naive T-cells divide on average once every 3.5 years. Secondly, the conclusion that cell survival and division are independent seems to be based on observations of T and B-cells activated by combinations of IL-2/αCD3 and IL-4/αCD40 respectively[5]. In these studies it was observed that the rate of cell death remained constant in the period between cell activation and

[3] Figures for the age range 0-20 years. As CD4+ T-cells comprise approximately two-thirds of the thymic export (Kindt et al, 2007), the total thymic export is estimated at between $3x10^8$ and $2.1x10^9$ cells per day.

[4] See later subsection "Potential distributions" for a description of the log-normal distribution

[5] The T-cell study (Deenick et al, 2003) referred to in Hawkins et al is examining the IL-2 regulation in T-cells and therefore IL-2 is present in almost all the experiments

first division. This T-cell activation is the triggering of a naive T-cell to proliferate and differentiate into effector and memory cells in response to stimulation by foreign antigen, with IL-2 considered to favour the production of effector cells (Surh and Sprent, 2008). As discussed above, homeostatic T-cell division appears to be driven by IL-7 and TCR-specific self-antigen, the same signals required for T-cell survival, and not IL-2. Therefore, while this study may provide evidence that T-cell survival and *activation* may be independent, it does not support the conclusion that T-cell survival and homeostatic proliferation are independent. In contrast the data from Bains et al supports the opposite conclusion, that homeostatic survival and proliferation *are* dependent, as the evolution of cell loss and division rates over time are closely correlated.

The stochastic niche model

Stirk et al (2008) propose an interesting model for T-cell homeostasis. The theory is that the immune system does not have the capacity to contain T-cells with TCRs that are exactly specific to every possible foreign antigen. Therefore optimal antigen coverage is provided by a repertoire of T-cells which react to a range of antigens but also have T-cells sufficiently diverse that they have minimal overlap in the antigens they interact with. It should be noted that, regardless of the level of specificity of TCRs to foreign antigens, low specificity interactions between TCRs and self-antigens are necessary for homeostatic T-cell survival and proliferation. Given the low specificity of these homeostatic interactions, it is likely that TCR of each clonotype interacts with a variety of self-antigens. Therefore this model is likely be useful in exploring T-cell homeostasis.

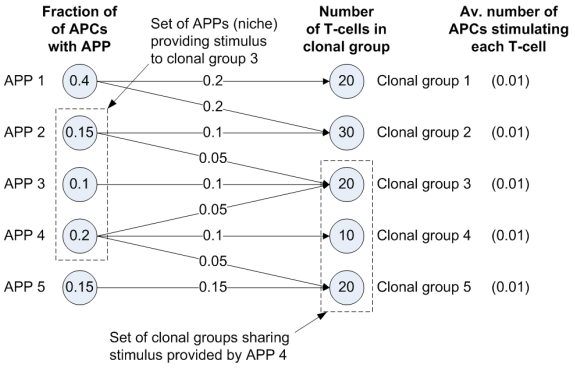
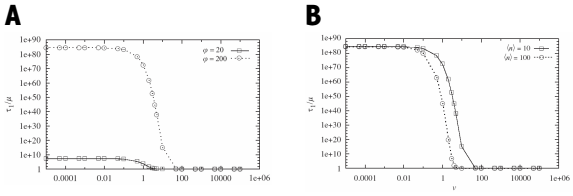


Figure 2: Toy illustration of the distribution of survival stimulation across clone members in a stochastic niche model. An APC is an antigen-presenting cell. An APP is an antigen presentation profile, representing a particular combination of antigens presented by an APC. In this toy example, there are only 5 distinct APPs. However, in the real immune system there will be many more APPs, each with very few APCs displaying that profile. In fact, given the vast number of potential antigen combinations and the fact APCs change their APP over time, it is likely that most APPs will be displayed by less than 1 APC on average. Note that, despite large differences in the fraction of APCs displaying each APP and the size of each clonal group, on average each T-cell receives 0.01 APC worth of stimulation and the system is in equilibrium. *[Source: Adapted from Stirk et al, 2008]*

In this model all antigen presenting cells (APCs) display a range of different self-antigens, and each particular grouping of self-antigens is christened an *antigen presenting profile* (APP). The TCR of each clonotype can interact with a range of self-antigens, and therefore cells of each clonotype can receive survival stimulation from a range of APCs. Specifically, each member of a clonotype can receive a survival stimulus from any APC with an APP that includes one of the self-antigens with which the clonotype's TCR interacts. Where multiple clonotypes have TCRs which interact with at least one of the self-antigens in an APP, the model divides the survival stimulation from all the APCs in that APP evenly among all the members of those clonotypes. A simple illustration of this model is shown in figure 2. The set of APPs containing self-antigens with which a given clonotype's TCR can interact is its *niche*.

The model assumes a uniform mean T-cell death rate, and a T-cell birth rate for each clonotype that is dependent on the per-cell survival stimulus received by its members. The model then represents the evolution of the clonotype populations as a multi-dimensional Markov chain, before making some mean-field assumptions to separate the evolution of each clonotype as an independent one-dimensional Markov chain. Analysing one of these chains for a single clonotype, the authors derive the expected lifetime of a clonotype as a function of the clonotype's *niche overlap*[6] (ν_i); the clonotype's mean rate of receipt of survival stimuli (φ_i); the global mean T-cell death rate (μ); and the global mean clonotype size ($\langle n \rangle$). The key conclusion from their analysis is that clonotypes with $\nu_i \ll 1$ will have lifetimes many orders of magnitude larger than clonotypes with $\nu_i \gg 1$. Thus clonotypes with $\nu_i \ll 1$ will tend to out-live clonotypes with $\nu_i \gg 1$, driving the global average niche overlap down and maintaining a diverse repertoire of TCRs which covers all the available self-antigens.

The authors illustrate how the distribution of clonotype lifetime (τ_1) as a function of ν_i is affected by varying φ_i and $\langle n \rangle$ (see figure 3a-b). However, there are no good estimates for the distribution of φ_i, making it impossible to quantitatively estimate the expected clonal lifetimes. It is quite possible that φ_i and ν_i are not independent, as it is reasonable to expect that the mean time between survival stimuli (φ_i) may depend on the level of competition for APPs, which is measured by ν_i.



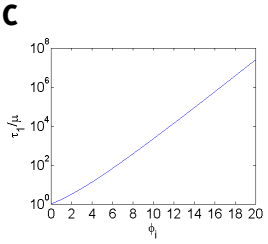


Figure 3: A-B: Effect of varying survival stimulation rate, φ_i (**A**) and mean clone size $\langle n \rangle$ (**B**) on the dependence of clonotype lifetime, τ_1 on clonotype niche overlap, ν_i. **C:** Effect of varying φ_i on the value of τ_1 for clonotypes with $\nu_i \ll 1$. In all plots τ_1 is expressed as a multiple of μ. *[Source: Stirk et al (2008). Figure C is generated from equation 31]*

[6] Niche overlap is the mean number of competing clonotypes for a given APP.

The authors make a point of stating that their model predicts that all clonotypes are doomed to extinction. Whether this is practically true is strongly dependent on φ_i. With estimates for the lifetime of naive T-cells (μ) ranging from 123-1811 days (Borghans and de Boer, 2007), clonotypes with $v_i \ll 1$ will outlive their human host (τ_1=128 years) so long as each T-cell of that clonotype receives a survival signal at least 8 times per day (φ_i=8). However, if φ_i<1.5, the lifetime of these clonotypes will be less than twice the mean lifetime of a naive T-cell (see figure 3c). While this still equates to clonotype lifetimes of 5-10 years for the upper estimates of μ, it is clear that, in this case, all clonotypes will be expected to become extinct over the lifetime of the host.

An interesting future study would be to use the model to simulate the generation of the naive peripheral T-cell population from scratch and compare the resultant age-dependent proliferation and loss to that observed in nature.

The data

The data used to drive and evaluate the model explored in this case essay was taken from two analyses which combine data on T-cell concentration, body mass, blood volume, TREC numbers and Ki67 expression to estimate the total population, thymic export, division rate and loss rate for naive $CD4^+$ cells (Bains et al, 2009; Bains et al, in press).

Raw data from the Bains et al analyses were supplied by the authors, providing estimates for thymic export, division rate and loss rate for ages 0-20 years, at intervals of 1 month. This data was interpolated[7] to provide daily estimates of these parameters. The daily division, loss, total population, net gain/loss and mean expected lifetime were then derived from this interpolated data by the following method and are illustrated in figure 4.

1. Initial population = previous day's total population + today's thymic export
2. Total population = initial population + initial population * (today's division rate - today's loss rate)
3. Net gain or loss = Today's total population - previous day's total population
4. Mean expected lifetime = 1 / today's loss rate

The first interesting feature to note is that the division and loss rates estimated by Bains et al exhibit essentially exponential decay. The sole small deviation is that the loss rate displays an initial stage of non-exponential growth before settling into exponential decay. This initial period of increasing loss rate corresponds to the first 168 days of the establishment of the peripheral T-cell population, during which the total population of naive CD4+ cells increases from 0 to 5.5×10^{10}. Given the requirement by T-cells for survival signals discussed previously, it might be hypothesised that this initial increase in loss rate corresponds to the increase in competition for survival signals to some constant equilibrium. This would be consistent with the stochastic niche model.

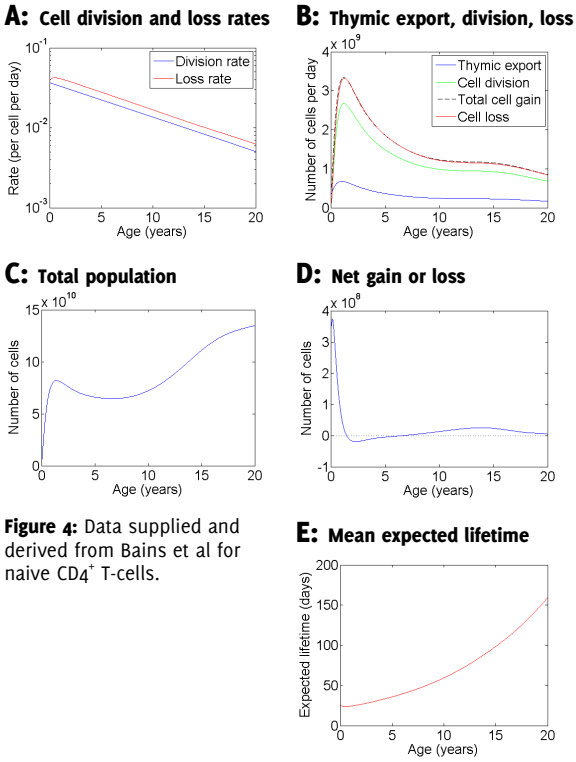


Figure 4: Data supplied and derived from Bains et al for naive $CD4^+$ T-cells.

Interestingly, the division rate exhibits no such initial increase, instead maintaining a constant exponential decline from day 1. This suggests that the level of any competition for resources modulating the division rate is constant from day 1 and does not increase with the total volume of cells. However, following the initial non-exponential growth period for the loss rate, both division and loss rates decay exponentially with the same time constant. This implies that, while there may be some differences in their modulation at low absolute cell numbers, their modulation is unlikely to be independent. This is consistent with the fact that IL-7 and self-antigens are both required for T-cell survival and can stimulate T-cell proliferation (Surh and Sprent, 2008).

The second interesting thing to note is that the total daily gain in cell numbers from thymic export and cell division appears extremely close to the daily cell loss. Indeed, on initial examination of figure 4b it would be tempting to conclude that they were so nearly identical that the total cell population would remain approximately constant with age. However, the small differences in total cell gain and loss represent daily net gains or losses of $\sim 10^7$-10^8 cells, as shown in figure 4d. This plot of net gain/loss clearly illustrates the significant net daily gain in the early years of peripheral development which drives the initial steep rise in total population. Later, much smaller net losses and gains (of the order of 2% of their gross counterparts) drive the subsequent gentler fall and further rise of the total population. The key point to draw from this is that small differences in the gross gain or loss curves can result in a significantly different population profile.

Finally, figure 4e illustrates the exponential increase in mean expected lifetime from a minimum of 24 days at 6 months to a maximum of 159 at 20 years. This is estimated by taking the inverse of the daily cell division rate. It is this trend that the static thymic distribution model seeks to explain.

[7] Cubic spline interpolation was performed using the interp1 function in Matlab (The Mathworks, 2007). The raw data for the first and final months in the range did not include an estimate for loss rate. Therefore daily loss rate estimates for the first and final months were extrapolated using the same method.

The static thymic distribution model

This is an attempt to explain the increase in the mean expected lifetime of peripheral naive T-cells with the age of the host using a very simple model. The fundamental premise of the model is that a suitable fixed distribution of expected lifetimes for T-cells exiting the thymus can result in a time-varying distribution of expected lifetimes for the peripheral population. Specifically, a distribution with most of its mass concentrated near the origin should result in the majority of each day's thymic output dying within days or weeks of export. If this distribution also possesses a long tail, then a small number of cells from each day's thymic export would have expected lifetimes measured in months or years. These long-lived survivors would accumulate over time as the peripheral population developed, gradually shifting the mean expected lifetime away from the mean of the daily thymic export and towards its tail.

Biologically, such a distribution of lifetimes for thymic emigrants might arise as the result of continued selection pressure in the periphery - if the "fitness" of thymic emigrants for survival in the periphery was similarly distributed. Given that $\sim 10^8$ new T-cells are exported from the thymus each day (Bains et al, in press) and there are only estimated to be $\sim 10^7$ distinct T-cell clonotypes in the periphery (Arstila et al, 1999) it seems reasonable to presume that the vast majority of each days thymic export will fail to compete successfully to establish a new clonotype and die. One potential flaw with this line of reasoning is an implicit assumption that all cell loss will be via cell death. Another alternative path for naive cell loss is conversion to a memory cell type. Data on the rate of conversion of naive cells to memory cells would be needed to determine if the number of cells lost via this path daily would be sufficient to "make room" for a significant proportion of the daily thymic export.

Potential distributions

Four different distributions were explored to assess their suitability for modelling the lifetimes of thymic emigrants.

Normal

The normal distribution is used to model the distribution of a wide range of properties for a variety of populations. However, it is not suitable for modelling the low mean, long tailed lifetime distribution required for this model. Due to its symmetric nature, a normal distribution with a mean of 24^8 and a tail extending for a significant number of days would have a significant proportion of its mass left of the origin (see figure 5a). As lifetime cannot be negative, this will severely distort the distribution. For completeness, the properties of a range of narrow normal distributions (with effectively all of their mass right of the origin) were explored. However, for all these distributions, the mean population lifetime rapidly decreased with time to be *less* than that of the thymic emigrants.

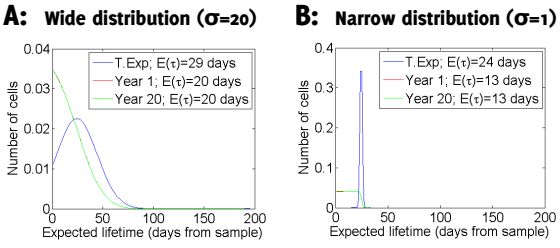


Figure 5: Evolution of population lifetimes where the export lifetimes of thymic emigrants are normally distributed. T.Exp shows the distribution of lifetimes for thymic emigrants. Years 1 and 20 show the distribution of lifetimes in the population as the periphery develops. These overlap for both plots.

Log-Normal

Limpert et al (2001) suggest that, despite the widespread use of the normal distribution, life is often actually log-normal. A log-normally distributed variable with parameters μ and σ is one whose logarithm is normally distributed with parameters μ and σ. The parameter μ controls the shape of the distribution, which morphs from a near-normal distribution for very low values of μ through increasingly left-skewed distributions to an "exponential-like" distribution at higher values of μ.

It is known that the normal distribution is an appropriate model for the likely spread of variable values if the source of variation is the *additive* combination of many independent effects. As a product can be rewritten as a sum of logs, it follows that the log-normal distribution is an appropriate model for the likely spread of variable values when the source of variation is the *multiplicative* combination of many independent effects.

It is not clear how the distribution of expected T-cell lifetimes could be considered to be due to such multiplicative variation. However, Hawkins et al state that their measured distribution of time-to-die for in-vitro B-cells is well fitted by a log-normal distribution. However, it should be noted that Hawkins et al also found the distribution of time-to-die is well fitted by the gamma and Weibull distributions. Sample log-normal distributions are shown in figure 6a.

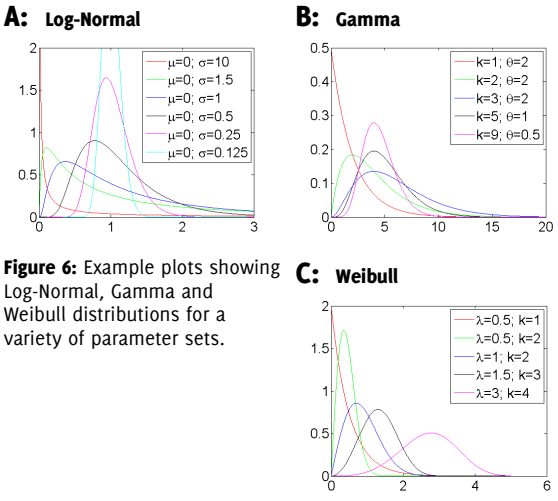


Figure 6: Example plots showing Log-Normal, Gamma and Weibull distributions for a variety of parameter sets.

Gamma

The gamma distribution is defined in terms of a shape parameter k and a scale parameter θ. The distribution morphs from "exponential-like" at $k \leq 1$, through decreasing left-skewed distributions, to a near-normal distribution at higher values of k. When k is an integer, the distribution is equivalent to the

[8] The minimum lifetime from the Bains et al data.

sum of k independent exponentially distributed variables, each with time constant $\tau=\theta$. When k=1, the gamma distribution is equivalent to the exponential function.

The "sum of exponential variables" equivalence means that a gamma distribution would be a good for modelling the expected time for an event to occur k times - if the event had a constant probability per unit time and thus the time between events was exponentially distributed. Therefore, this distribution might be a good model for cell loss if peripheral T-cells died due to consecutive shortening of their telomeres[9] with each cell division, and the interval between divisions was exponentially distributed. However, it is unlikely that this is the case. Measurements of telomere shortening in naive T-cells suggest that they divide on average 14 times prior to becoming memory cells (Borghans and de Boer, 2007). If telomeres can divide 14 times prior to becoming memory cells, they must divide at least this many times before they die due to telomere loss. Inspection of the sample gamma distributions in figure 6b makes it clear that such a gamma distribution would be a "normal-like" distribution. This distribution would not be left-skewed and could not simultaneously concentrate its mass near the origin *and* have a long tail.

Weibull

The Weibull distribution is defined in terms of a shape parameter k and a scale parameter λ. It is widely used in the analysis of mean time between failures, with the value of k describing how the failure rate changes with time. If k=1 the failure rate is constant over time and, like the gamma distribution, the Weibull distribution becomes equivalent to an exponential distribution with time constant $\tau= \lambda$. If k>1, the failure rate increases with time, and older members of the population are more likely to fail ("wearing out"). If k<1, the failure rate decreases with time and defective items fail at a young age ("infant mortality"). This latter case describes precisely the premise of the model. Therefore the Weibull distribution appears to be more transparently aligned to the model requirements than the other distributions. Sample Weibull distributions are shown in figure 6c.

Model design

The model was constructed to use the daily estimates for total thymic export and cell division derived from the Bain et al data as its input. The initial model used for distribution selection ignores cell division, and starts with an empty periphery containing no cells. Each day it adds the appropriate number of thymic emigrants according to the Bains et al data, but gives them a distribution of lifetimes according to the selected distribution. Each day, following the addition of new cells from the thymus, all cells with an expected lifetime of zero days are removed from the population, and the expected lifetime of all the remaining cells is reduced by one day.

Following distribution selection, cell division was incorporated into the model in order to put the population profile predicted by the model on the same scale as that derived from the Bains et al data. This raised the question of which cells are dividing. It could be reasonably hypothesised that older cells, which have been more successful at dividing, are also more successful at proliferation. This hypothesis is supported by the fact that cell survival and cell division rely on some of the same stimuli, as well as by the correlation between changes in cell loss and division rates observed in the Bains et al data. However, modelling such differential division requires the selection of a model for this differentiation, along with the estimation of its associated parameters. In addition to the fact that the incorporation of a cell differentiation model would make the overall model more complex, it is likely that there is insufficient independent data to sufficiently constrain the additional parameters. Thus, fitting the data would not necessarily permit any conclusions to be drawn about the biological plausibility of the model. Therefore, for this case essay, the daily thymic output was simply "topped up" by the number of new cells produced by cell division, giving these new cells the same distribution of lifetimes as the thymic emigrants. This keeps the model simple. As the daily number of new cells due to cell division in the Bains et al data is very close to four times that day's thymic output, it would also seem that such topping up well approximates the situation were all cells divide an equal number of times in their lifetime (in this case ~4). However, it is recognised that such a situation is only compatible with this model's heavily left-skewed distribution of lifetimes if all this division occurs on the day of a cell's output from the thymus. This is clearly biologically implausible as, even when being artificially driven to divide in vitro, B-cells take from 30-50 hours to enter into their first division cycle and then ~9 hours to complete subsequent division cycles (Hawkins et al, 2007). It is likely that maximum proliferation rates for T-cells are similar. Yet, even if such rapid division was plausible in vivo, four divisions would take a minimum of 2.5 days. The impact of various models for cell division is discussed further later in this case essay. The model's algorithm, incorporating cell division, is summarised below.

Model algorithm

1. Take surviving cells from previous day and add thymic export and new cells produced by division according to Bains et al data, with lifetimes distributed according to selected distribution.
2. Remove all cells with an expected lifetime of zero. This will include a significant proportion of today's thymic export and new cells produced by division.
3. Reduce expected lifetime of all remaining cells by 1 day.

Distribution selection

It would appear that any one of the log-normal, gamma or Weibull distributions is be capable of producing the heavily left-skewed, long-tailed distributions required for the model. However the Weibull distribution was favoured as it's parameterisation most closely matched the premise of the model. This was due to having a parameter that could be used to explicitly vary the distribution of loss between young and old peripheral T-cells. However, for the sake of completeness, an initial parameter exploration was carried out for each of the

[9] Telomeres are protective caps consisting of repeating DNA segments at the end of chromosomes. DNA replication cannot continue right to the end of the chromosome during cell division, so a segment of telomere is lost during each division.

log-normal, gamma and Weibull distributions. In this initial parameter search, a variety of parameter sets were chosen for each model to cover the range of distribution shapes they were capable of producing. In this initial exploration, the model did not include any cell division. The parameters selected for each distribution in this initial exploration were restricted such that the analytically derived mean lifetime was 24. The parameter sets used are listed in the appendix, and the most extreme left-shifted distributions of each type are illustrated in figure 7.

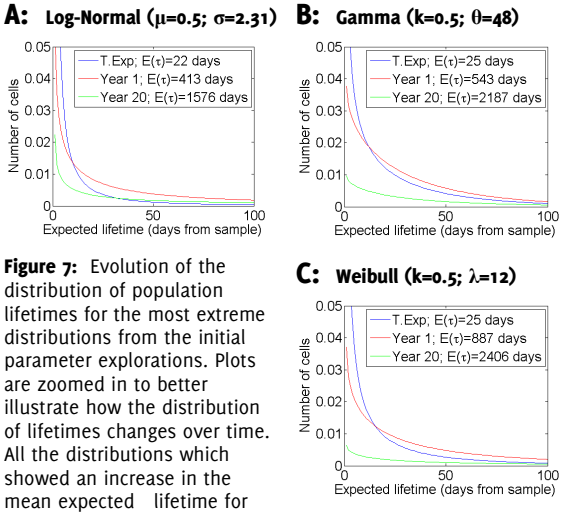


Figure 7: Evolution of the distribution of population lifetimes for the most extreme distributions from the initial parameter explorations. Plots are zoomed in to better illustrate how the distribution of lifetimes changes over time. All the distributions which showed an increase in the mean expected lifetime for the population were "exponential-like" curves, with very steep gradients near the origin that later flattened to very shallow gradients to provide a long tail. Note that these are the original, *incorrect* population lifetime distributions used for model selection. The corresponding corrected distributions are shown in figure 8. However, the observation that all suitable distributions are "exponential-like" still holds true.

These represent the limit of what was possible with the model at this time. The more extreme the left-shift of the distribution, the more extreme the concentration of mass near the origin. In order to sample such distributions accurately for the model, higher sampling frequencies have to be used. As the distributions become extremely left shifted, the sampling frequency grows to the point that there is insufficient memory to hold the distribution. This problem was substantially alleviated in later versions of the model, with the introduction of variable sampling frequency across the distribution, permitting concentration of mass near the origin to be selectively sampled at much higher rates.

A crucial observation is that all the distributions that showed an increase in the mean expected lifetime for the population were "exponential-like" curves, with very steep gradients near the origin that later flattened to very shallow gradients to provide a long tail. This was not expected. It was initially thought that any substantially left-skewed distribution would suffice. This finding can be understood by considering the Weibull distribution, where all "infant mortality" curves with k<1 are of this form. This might reasonably be interpreted as meaning that *any* distribution which results in a high early loss rate and a low later loss rate must be of the same form.

The most extreme left-skewed distributions for all three classes resulted in increases in mean expected lifetimes well in excess of that observed in the Bans et al data (159 days). As the parameterisation of the Weibull distribution allowed direct manipulation of the distribution of loss between young and old peripheral T-cells, it was selected for further investigation.

Problems with the original distribution selection

During the further investigation of the model, the model developed from the version used in distribution selection. One of the key developments was the identification of an artefact introduced by the original method of sample normalisation (required to handle the inevitable "missing" mass in a finite discrete sample of a continuous distribution). This is discussed in detail in the appendix. At the time the normalisation error was discovered and rectified, it was not thought to check for any impact on distribution selection. The full evaluation of the model was therefore performed using the Weibull distribution.

However, this error did transpire to have a significant impact on distribution selection. The error caused the distributions of population expected lifetimes used for selection were heavily skewed towards high lifetimes. These distributions were recreated using a more appropriate method of sample normalisation and are shown in figure 8. It should be noted that the mean expected lifetimes calculated for the original thymic export curves plotted in figure 7 were also found to have been calculated incorrectly. They were calculated from the distribution *prior* to the addition of the missing mass to the final day. If this error had not been made, it would have been clear from the mean lifetimes of the distributions that the sampling was not appropriate, as the sample mean lifetimes would have been significantly higher than the (analytically calculated) expected lifetime of 24.

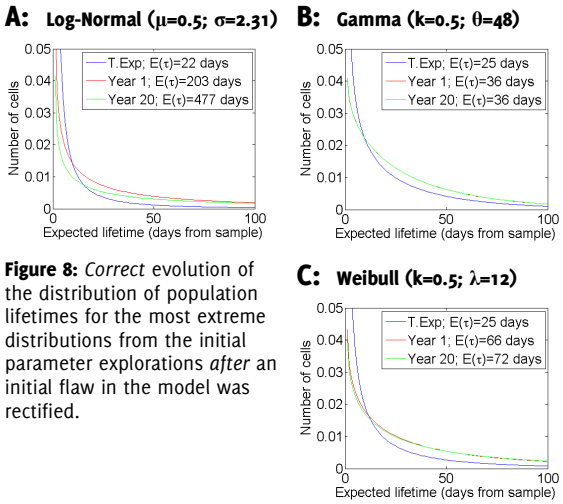


Figure 8: *Correct* evolution of the distribution of population lifetimes for the most extreme distributions from the initial parameter explorations *after* an initial flaw in the model was rectified.

Examining the corrected distributions it is clear that, despite the parameterisation of the Weibull distribution seeming a closer match to the model, the log-normal distribution appears to more closely replicate the desired behaviour. Mean lifetimes for both the gamma and Weibull distributions approach their maximum at early host ages, while the mean lifetime for the log-normal continues to evolve until much later. There was not sufficient time to perform a robust exploration of the log-normal parameter space. However, a limited ad-hoc exploration suggests that this distribution may be no better at explaining the data than the Weibull distribution.

Model evaluation

Parameter exploration

Following the selection of the Weibull distribution for thymic export, a wide range of parameter sets were explored. The "fitness" of each parameter set was evaluated by selecting two "evaluation" variables from amongst the model's output. The two variables chosen were daily population and daily loss. These were selected as they were the only two independent outputs[10], with the remaining outputs being a function of one or both of these. Two separate error surfaces were constructed by calculating the root mean squared (RMS) error over the profile of each variable for a range of different parameter sets.

In total 120 different parameter sets were explored. First a wide range of the parameter space was sampled at a relatively coarse resolution. This exploration was limited to values of k<1, as analysis of the Weibull distribution during model selection had already established that the mean population lifetime only decreased with host age for values of k below 1. Once the general shape of the error surfaces for each of the two evaluation variables had been established, a finer sampling of the parameter space was made around the minimum error region of the error surfaces. A full list of the explored parameter combinations is included in the appendix, and the final error surfaces are shown in figure 9.

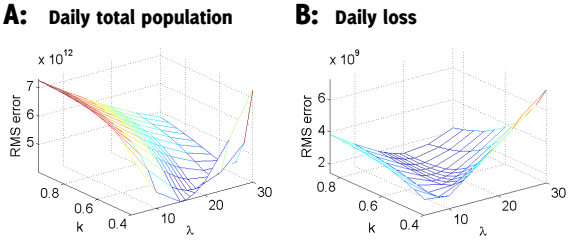


Figure 9: Error surfaces for the two evaluation variables, showing how the RMS error for each of the two evaluation variables. Note that 0.4 represents the lower limit of k that can be explored using the current minimum variable sampling frequency of 10^{-6} days.

Analysis of "best" parameters

Examining the error surfaces for the two evaluation variables reveals that the minima for the two surfaces arise for different parameter sets (λ=15; k=0.4 for total population vs. λ=20; k=0.8 for daily loss).

Minimum total population RMS error

The value of k that produces the minimum error for the population profile is 0.4. This is at the edge of the explored parameter space, and it is therefore possible that a lower minimum error might be achieved with a lower value for k. However 0.4 represents the lower limit of k that can be explored using the current maximum sampling frequency of 10^{-6} days.

Changing the model to permit a higher maximum sample frequency is possible, but there was not time to make this improvement. Examining the model output for this "best" parameter set (see figure 10), it can be seen that the shape of the population profile with minimum RMS error is qualitatively different to the profile derived from the data, lacking the secondary rise observed to commence at a host age of ~7 years. The daily loss rate and mean lifetime profiles are also poor fits to the data.

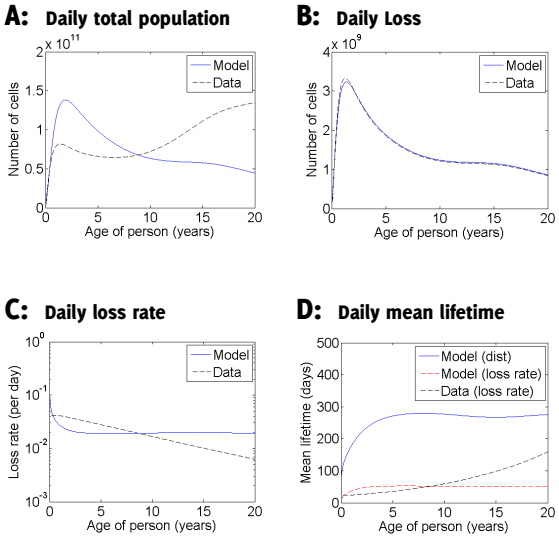


Figure 10: Key model outputs compared with the data from Bains et al. These outputs are for the "best" model parameters as defined by minimum RMS error for total population (λ =15; k=0.4). Note the semi-log axis on for the daily loss rate plot.

Best subjective fit

The outputs of the model for all 120 explored parameter sets were visually inspected, and it was observed that the qualitative shape of the population profile was well conserved across the explored parameter space. A subjective assessment of the "best" fit to the data was therefore made, considering this general difference in profile shape. The model outputs for this subjective "best" parameter set are shown in figure 11. This subjective best fit is clearly unlikely to be approached by minimising the RMS error of the population profile. Thus the impact of not exploring the parameter space for k<0.4 on the validity of the conclusions drawn in this case essay is likely to be small.

[10] The mean lifetime as estimated from the distribution of population lifetimes is not a function of either daily population or daily loss, but this is not derivable from the Bain et al data, so cannot be used to evaluate the model.

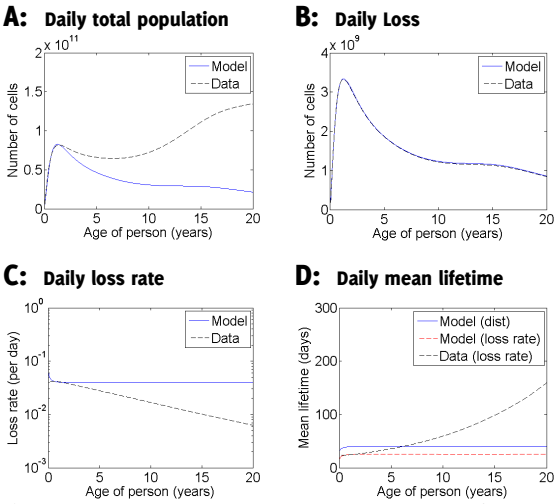


Figure 11: Key model outputs compared with the data from Bains et al. These outputs are for the parameter set subjectively judged the best fit to the data after visual inspection (λ =20; k=0.7). The outputs for this parameter set are very similar to those producing the minimum RMS error for daily loss. Note the semi-log axis on for the daily loss rate plot.

Minimum daily loss RMS error

The "best" model outputs for the minimum error in the daily loss profile are close to those subjectively judged "best" (see figure 12. This suggests that daily loss is an appropriate model evaluation parameter. During the following analysis, unless otherwise specified, all discussion refers to the model output for this parameter set.

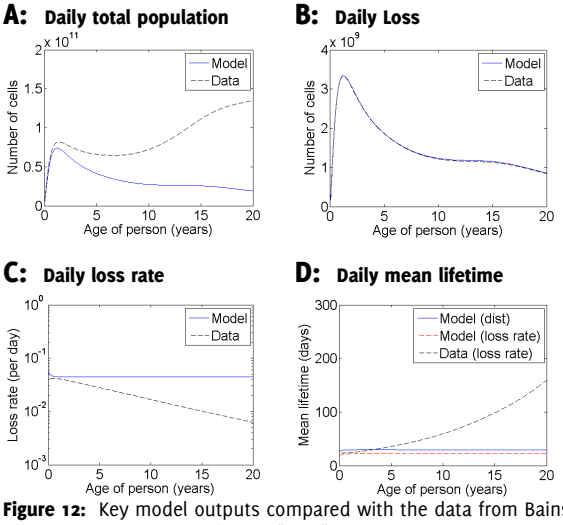


Figure 12: Key model outputs compared with the data from Bains et al. These outputs are for the "best" model parameters as defined by minimum RMS error for daily loss (λ =20; k=0.8). Note the semi-log axis on for the daily loss rate plot.

Further analysis

None of the "best" fit parameters result in an especially good fit to the data for any of the model outputs. The fit of the model daily loss profile to the data appears on first inspection to be very good. However, very small differences between the model and data profiles can translate into differences of $\sim 10^7$ between the daily net gain or loss calculated by the model and that derived from the data. Figure 13 shows that, even for the parameter set that minimised the error in the daily loss, such large differences in daily net gain/loss still occur.

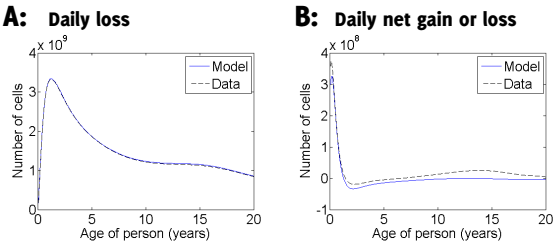


Figure 13: Illustration of how differences in daily loss almost imperceptible by visual inspection can cause large differences in the daily net gain or loss between the model and the data

There are two key differences between the model outputs and those derived from the data. The first is a substantially diminished model population following the initial peak at year 1. The data-derived population gently falls between the ages of 1 and 6, before gradually rising again up to the age of 20, where it starts to plateau. In contrast, the model-derived population profile mirrors the shape of the thymic export and division profiles that form the input to the model (see figure 4b), falling continuously in an exponential-like decay between the ages of 1 to 15, before increasing its rate of decline again. It is hypothesised that the difference in population profiles is due to a difference in the evolving distribution of lifetimes between the model and the data. As can be observed in figure 12d, the mean population lifetime derived from the loss rate in the data increases exponentially with age, while that derived from the loss rate in the model rapidly converges to a fixed value. As these derived population lifetimes are simply the inverse of the daily loss rates, further explanation will be couched in these terms. Interestingly, the mean lifetime of the population calculated directly from the model lifetime distribution was greater than that calculated from the model loss rate for all the parameter sets explored. The difference between these two estimates shrinks as k increases towards 1. This difference likely reflects the fact that the inverse loss rate is only robust estimate for mean lifetime for true exponential distributions, which have a constant loss rate with time. For Weibull distributions with k<1, the loss rate decreases with time, resulting in an underestimate of the mean lifetime from the loss rate.

The loss rate derived from the data shows an initial rise to the age of 6 months, before settling into an exponential decay from the age of 1 year onwards. In contrast, the model loss rate experiences a steep fall from an initially high rate, levelling out to a constant rate of loss from the age of 1 year onwards. The difference between the behaviour of the distributions in the first year is easily explained by the fact that year 1 is when the peripheral population peaks. The peripheral population start off at zero, gradually rising to a maximum at the age of 1. This means that, in vivo, competition for survival resources in the periphery will start very low and rise to a maximum as the population peaks. This accounts for the rising death rate during the first year observed in the data. In the model, there is no concept of competition for survival resources. With the "topping up" approach used in the model, the distribution of lifetimes for cells produced by division is the same as that for thymic emigrants. With an empty periphery, the loss rate for the first day is simply the ~9% of the

new cells with an expected lifetime of zero days. However, on subsequent days a small but increasing proportion of the total population is composed of "survivors" from previous thymic exports with longer than average lifetimes. This gradually lowers the loss rate of the population as a whole, eventually reaching an equilibrium rate at 1 year.

A more difficult question is why the model loss rate tends to a fixed equilibrium, while the data loss rate tends to an exponential decay. The decreasing loss rate for the data is compatible with preferential division by cells with long expected lifetimes. Assuming that a significant proportion of the dividing cells live long enough to divide many times, these long-lived cells will be *exponentially* increasing in number. In contrast, the numbers of short-lived cells that divide few times will be added to primarily by the *linear* addition of cells from the thymus. The exponential expansion of the long-lived cells will rapidly dominate the linear expansion of the short-lived cells, resulting in an exponential increase in the population mean lifetime and therefore an exponential decrease in loss rate.

As discussed in the "Model design" section, due to the close correlation between the daily numbers of thymic emigrants and new cells produced by division, the "topping up" model for division might be considered equivalent to a scenario where each cell divides the same number of times in its lifetime. With no preferential increase in the number of long-lived cells, it is not surprising that the population mean lifetime does not drift far from the mean lifetime for newly created cells.

Conclusions

It is clear that the model as it stands does not explain the data well. A static distribution of lifetimes for all newly generated cells is therefore not sufficient to explain the observed increase in the mean population lifetime with age. Even so, it may still be possible to explain this increase with a fixed distribution of lifetimes for thymic emigrants *only*. However, assuming this can be achieved, it is likely to require a model for cell division that results in the preferential proliferation of long-lived cells.

Differential cell division

One potential method of achieving such differential cell division might be to simply give the newly divided cells the same distribution of lifetimes as the whole population. This might be considered equivalent to giving each cell the same probability of division, with longer-lived cells simply having more opportunities to divide. This would be expected to significantly accelerate the accumulation of longer-lived survivors compared to the current method, but it is not clear if this would be sufficient to match the exponential increase in mean lifetime observed in the data. The division of cells could be further biased towards long-lived cells by incorporating the 30-50 hour time-to-first-division delay observed by Hawkins et al (2007). This would prevent each day's thymic emigrants from contributing to that day's division. As ~9% of thymic emigrants fail to survive to the day after they are created, this might be expected to have a further significant effect on the evolution of the population mean lifetime.

If neither of these methods are sufficient to replicate the observed increase in mean lifetime, one final possibility would be to explicitly give longer lived cells a greater probability of dividing. One way to do this might be to produce a new "division allocation" distribution, calculated by scaling the "raw" population lifetime distribution by some inverse function of the expected lifetime, and allocate the newly divided cells according to this.

Exploration of the log-normal distribution

Finally, there is a possibility that the log-normal distribution might be more appropriate for modelling the lifetime distribution for thymic emigrants. The basis for such speculation is the fact that mean lifetime using the log-normal distribution appears to continue to increase more quickly, and for a longer period of time, than is the case with the Weibull distribution. The final mean lifetime is also much bigger (see figure 7). For reasons that are explained in the "Model selection" section, this property of the log-normal distribution was not discovered until recently, and there has not been sufficient time to replicate the robust parameter exploration performed for the Weibull distribution.

A very limited ad-hoc exploration of selected log-normal distributions suggests that the model output is qualitatively similar to that produced using the Weibull distribution, and therefore it is possible that it will be no better at explaining the data than the Weibull distribution. However a fuller parameter exploration is recommended.

Acknowledgements

Thanks to Robin Callard and Iren Bains for their expertise and advice. Thanks also to Anton Flügge and Robert Henderson for interesting and useful discussions on the subject. Additional thanks to Iren Bains for providing data on thymic output and T-cell division and loss.

All modelling and analysis was done using Matlab (The Mathworks, 2007). Definitions of the various distributions explored were taken from Wikipedia (http://www.wikipedia.org) and confirmed at Wolfram MathWorld (http://mathworld.wolfram.com).

UCL CoMPLEX is an EPSRC-funded Life Sciences Interface Doctoral Training Centre.

References

Arstila, T. P.; Casrouge, A.; Baron, V.; Even, J.; Kanellopoulos, J. & Kourilsky, P. (1999). A Direct Estimate of the Human T Cell Receptor Diversity. Science *286*, 958-961.

Bains, I.; Antia, R.; Callard, R. & Yates, A. J. (2009). Quantifying the development of the peripheral naive CD4+ T cell pool in humans. Blood, pre-published online January 28, 2009. DOI 10.1182/blood-2008-10-184184

Bains, I.; Thiebaut, R.; Yates, A. J. & Callard, R (in press). Quantifying Thymic Export: combined models of naive T cell prolifereation and TREC concentration gives explicit measure of thymic output. In press.

Borghans, J. A. M. & de Boer, R. J. (2007). Quantification of T-cell dynamics: from telomeres to DNA labeling. Immunological Reviews *216*, 35-47.

Deenick, E. K.; Gett, A. V. & Hodgkin, P. D. (2003). Stochastic Model of T Cell Proliferation: A Calculus Revealing IL-2 Regulation of Precursor Frequencies, Cell Cycle Time, and Survival. J Immunol *170*, 4963-4972.

Hawkins, E. D.; Turner, M. L.; Dowling, M. R.; van Gend, C. & Hodgkin, P. D. (2007). A model of immune regulation as a consequence of randomized lymphocyte division and death times. Proceedings of the National Academy of Sciences *104*, 5032-5037.

Kindt, T.; Golsby, R.; Osbourne, B. & Kuby, J. (2007). Kuby Immunology (6ed). *W.H. Freeman and Company*.

Limpert, E.; Stahel, W. A. & Abbt, M. (2001). Log-normal Distributions across the Sciences: Keys and Clues. BioScience *51*, 341-352.

Stirk, E. R.; Molina-París, C. & van den Berg, H. A. (2008). Stochastic niche structure and diversity maintenance in the T cell repertoire. Journal of Theoretical Biology *255*, 237-249.

Surh, C. D. & Sprent, J. (2008). Homeostasis of Naive and Memory T Cells. Immunity *29*, 848-862.

The Mathworks (2007). Matlab 2007a. Natick, Massachusetts, USA.

Appendix

Sampling the thymic export PDF

Sampling frequency

For the initial model exploration used for distribution selection, the thymic export probability density function (PDF) was calculated in the range 0 to 20 years + 1 day, with the interval between data points (the sampling frequency) determined by the heuristic previously described in the "Distribution selection" section). This "oversampled" PDF was then summed over each day to produce a "histogram" PDF with one bin per day.

For the subsequent "coarse" parameter exploration using the Weibull distribution a uniform sampling interval of 0.001 days was used. Following this "coarse" exploration it became clear that, in order to more accurately determine the shape of the parameter "fitness space", it would be necessary to include parameter sets with $k<0.5$. Such parameter sets required a much higher sampling frequency in order to capture the concentration of mass near the origin. However, memory limitations meant that it was not possible to uniformly sample at a high enough frequency. Therefore, for the finer scale exploration of the Weibull distribution, a mixed sampling density was used, with the first day sampled at an interval of 10^{-6} days and the remaining days sampled at an interval of 0.001 days.

The minimum host age sampled was set at 0.001 (for the coarse exploration) or 10^{-6} days (for the fine exploration) in order to eliminate numerical explosions at zero. This was required as, with $k\leq1$, the Weibull distribution asymptotically approaches infinity at the origin. The coarse and fine sampling intervals were set to ensure the sampling of the underlying PDF was at sufficiently high resolution that, if the upper bound was set high enough, the area of the sampled PDF would exceed 0.999 and the sampled PDF mean would be within ±1 of the analytical PDF mean.

For consistency, the coarse level parameter exploration was repeated using the new variable sampling frequency and it is these repeated data that are included this report.

Sample normalisation

Initial model exploration for distribution selection

Any finite, discrete sampling of a continuous distribution will have some "missing mass". This problem was especially acute for the type of distributions required for this model, which were both heavily skewed to the left and very long tailed. Thus, in order to capture sufficient distribution mass for a good sample, they required *both* a high sampling frequency (to accurately capture the mass near the origin) *and* the sampling of the distribution to a high host age (to capture all of the mass in the long tail). However, memory constraints placed a practical limit on the maximum number of sample points.

In the early version of the model used for distribution selection, a fixed sampling rate was used. Consequently, simultaneously achieving a sufficiently high sampling rate and sampling to a sufficiently high host age required far more sample points than could be held in memory. During the initial model selection explorations, a heuristic method was therefore used to determine if the sampling rate was sufficiently high.

This heuristic involved restricting the sample to the 7306 days (20 years) required for the model and increasing the sampling frequency until there was no appreciable increase in the captured mass. Therefore it was assumed that any "missing" mass lay in the long tail beyond day 7306. This unsampled mass was typically ~1% of the total mass. Rather than re-normalising the truncated 7306 day sample, the missing mass was placed in day 7307, thus making the total mass of the distribution equal to 1. If the assumption that all the missing mass existed in the unsampled tail of the distribution was true, all this mass would have a lifetime of at least 7307 days. It was therefore felt that placing the missing mass in day 7307 would distort the sample mean lifetime less than distributing it evenly over days 1-7306 by re-normalising the sample. As the mass in this final day had a lifetime of 7307 days, it would not contribute to cell loss in the lifetime of the model (7306 days), so the precise distribution of lifetimes was irrelevant to the determination of daily population changes.

Subsequent model exploration

During the course of further investigation of the model using the Weibull distribution, it became desirable to sample this distribution at a higher frequency than permitted by memory constraints. In order to achieve this, the capability was developed to sample the distribution using a mixture of sampling frequencies (see above). This permitted the first day, containing the problematic concentration of mass near the origin to be sampled at a much higher rate. Using this mixed sampling frequency, it was possible to sample the distribution to a sufficiently high host age to establish that the bulk of the "missing mass" was not actually contained in the tail of the distribution, as was initially assumed. This meant that the missing mass was due to the finite sample rate and it would therefore be more appropriate to distribute it evenly across the sample by re-normalisation[11]. This meant that the distributions of population lifetimes used for the initial distribution selection were heavily skewed towards high lifetimes. This is discussed further in the main report in the "Distribution selection" section.

It is clear that the heuristic used to establish a sufficiently high sampling frequency during distribution selection was flawed. Examination of the sampled distributions generated during this process reveals that, for the "extreme" distributions, the sample area was very close to 1 (exceeding 0.999 in several cases). In retrospect, it is therefore likely that these distributions were assumed to be sampled at a sufficiently high frequency by virtue of their high area.

Parameter exploration

Model selection

The distributions and parameter sets used for model selection are shown in table A1. These parameter sets were selected in order to explore a

[11] Re-normalisation is achieved by dividing the value of each data point in the PDF by the area of the PDF.

range of parameter values for each distribution while keeping the mean of the thymic export PDF approximately 24 (to match the minimum mean lifetime observed in the data). All parameter values are accurate to four decimal places.

Weibull (n=12)	
λ	**k**
12	0.5
15.9513	0.6
18.96	0.7
21.1826	0.8
22.8097	0.9
24	1
25.2273	1.5
27.0811	2
26.8763	3
26.1390	4
26.5856	5
25.2273	10

Gamma (n=10)	
k	**θ**
0.50	48
0.75	32
1	24
1.50	16
1.75	13.7143
2	12
3	8
4	6
5	4.8
10	2.4

Log Normal (n=5)	
μ	**σ**
0.50	2.3143
0.75	2.2037
1	2.0871
2	1.525
3	0.5967

Normal (n=8)	
μ	**σ**
24	0.50
24	0.75
24	1
24	2
24	3
24	4
24	5
24	10

Table A1: Distributions and parameter sets used for initial model exploration

Further exploration using the Weibull distribution

The Weibull distribution was selected for further, more systematic, parameter exploration. Firstly, a coarse exploration of the parameter space was performed to ascertain the large-scale shape of the error surfaces. Subsequent finer scale parameter explorations were then performed to more accurately determine the shape of the error surfaces in the immediate area of the minimum. The parameter sets used were all combinations of the following values for λ and k, making a total of 120 parameter sets. Note that k=0.4 is the lowest value of k that can be explored with the current maximum variable sampling frequency of 10^{-6} days.

Parameters	
λ	**k**
5	0.40
10	0.45
15	0.50
16	0.55
17	0.60
18	0.65
19	0.70
20	0.75
25	0.80
30	0.85
-	0.90
-	0.95

Table A2: Distributions and parameter sets used for initial model exploration