CHAPTER 3

Development of the T cell repertoire in the thymus

Introduction

The repertoire of T cells generated by VDJ recombination in an organism exhibits some remarkable properties. T cells mount pathogen-specific responses to peptide-MHC molecules with peptides derived from a diverse and evolving world of microbes. At the same time, T cells can be cross-reactive to a few peptides. The T cell repertoire is largely self-tolerant, in that most T cells do not respond to most peptide-MHC molecules derived from endogenous proteins that are displayed on host cells. How does such a largely self-tolerant, pathogen-specific, yet somewhat degenerate, T cell repertoire develop? In this chapter, we consider some aspects of T cell development that shed light on this question. Immature T cells exit the bone marrow and then undergo development (or maturation) in an organ called the thymus, which is located behind the sternum (Fig. 1). The processes that occur in the thymus are complex, and detailed descriptions of the biology are available in standard immunology textbooks []. In this chapter, we will consider only some key aspects of these processes that shape some features of the T cell repertoire. We will then focus on how insights in to the consequences of these processes on the nature and function of the T cell repertoire can be obtained by complementary theoretical analyses, computation, experimental studies in mice, and clinical observations. Finally, we will end by noting the many outstanding questions that need to be studied in future.

Basic biology of thymic development

The importance of the thymus for maintaining a healthy immune system was recognized by the observation that removal of the thymus at birth in mice led to immunodeficiency. The thymus can be roughly partitioned in to an outer region called the cortex, and an inner region called the medulla, and it is comprised of a network of cells called the thymic stroma. In addition, the thymus also contains dendritic cells and macrophages. Precursors of T cells, called thymocytes, arrive in the thymus from the bone marrow. At this stage, the TCR genes have not rearranged, and these early stage thymocytes are called double negative thymocytes because they do not express the co-receptors, CD4 and CD8. Interactions with thymic stromal cells then result in various stages of thymocyte differentiation. During these differentiation steps, various surface molecules are expressed, the TCR β chain genes rearrange, CD4 and CD8 are both expressed, and then the TCR α chain also rearranges. At this stage, thymocytes are called double positive (DP) thymocytes because they express both types of co-receptors. Then, several stages of differentiation occur during which either CD4 or CD8 ceases to be expressed on the T cell, and two processes called positive and negative selection occur. Ultimately, a fraction of T cells that successfully undergo these processes exit the thymus and become part of the mature T cell repertoire that circulates in blood and peripheral tissues. In mice, the time duration from when an immature T cell precursor enters the thymus to when it departs as a mature T cell is about 3 weeks.

Usually, genes in multicellular organisms are expressed in a tissue-specific manner. The action of a gene called AIRE causes promiscuous expression of diverse host genes in thymic cells. Thus, thymic cells

express peptide-MHC molecules where the peptides are derived from diverse parts of the proteome. Thymocytes migrate through the thymus and interact with these self peptide-MHC molecules. For a thymocyte to successfully mature in to a peripheral T cell, it must pass two tests. It must bind to at least one of the self peptide-MHC molecules it encounters with an affinity that exceeds a threshold value (E_p). This process is called positive selection, and there is some evidence that this is mediated largely by peptide-MHC molecules presented on thymic epithelial cells. Thymocytes must also not bind to any of the encountered self peptide-MHC molecules with an affinity that exceeds another affinity threshold (E_n), which corresponds to stronger binding than E_p. This process is called negative selection, and many cell types, including thymic dendritic cells and macrophages play a role in mediating negative selection.

A set of experiments in mice provided insights into the positive and negative selection thresholds, E_p and E_n. Transgenic mice are genetically engineered animals whose T cells only express a single type of TCR. Such mice with a particular TCR, called OT-1, have been used extensively in experimental studies of diverse aspects of CD8 T cells. This TCR recognizes a particular peptide from the chicken ovalbumin protein, and the affinities and potencies of different variants of this so-called OVA peptide for the OT-1 TCR have been characterized. For example, the potency can be characterized by pulsing antigen presenting cells with the peptide so that it is presented on its surface. Then T cells or thymocytes can be stimulated by these antigen presenting cells, and a downstream marker of intracellular T cell, or thymocyte, signaling that serves as a proxy for activation can be measured. The variation of the extent of T cell signaling with peptide concentration exhibits a sigmoidal shape. A metric of the potency of a particular variant of the OVA peptide for stimulating OT-1 T cells or thymocytes is the peptide concentration that corresponds to half the maximum amount of downstream signaling. This quantity is referred to as EC_{50} (Fig. 2a). More potent peptides are characterized by a smaller value of EC_{50}. The affinities, or half-lives of binding, of the variant peptides for the OT-1 TCR can be assessed using surface plasmon resonance experiments. Peptides with higher affinities, or usually, longer half-lives of binding to the TCR (see Chapter 7), are more potent.

Palmer and co-workers characterized the nature of the negative and positive selection thresholds using the OT-1 system described above. Double positive thymocytes were taken from transgenic mice expressing the OT-1 TCR. The thymus of a fetal mouse can be cultured *in vitro* (fetal thymic organ culture or FTOC), and the number of single positive thymocytes that develop in FTOC pulsed with variant OVA peptides was assessed. The dependence of this quantity on the potency of the peptide (measured as described above) shows that the negative selection threshold is very sharply defined, while the positive selection threshold is somewhat softer (Fig. 2b). Comparison of these thresholds with measured OT-1 TCR affinities of the peptides that lie at the thresholds of negative and positive selection indicates that E_p and E_n are separated by a relatively small difference (a couple of thermal energy units, k_BT, where k_B is Boltzmann's constant and T is the absolute temperature at physiological conditions).

We close this brief summary of the biological processes that occur in the thymus during T cell development by noting that, in individuals past puberty, T cells develop in the thymus at low rates. Thymic output is not a big factor in maintaining the T cell repertoire in mature individuals. Rather, existing T cells replicate to maintain the repertoire.

The effect of thymic selection on the specificity of T cells for pathogenic peptides

Some years ago, a set of experiments carried out by Huseby and co-workers provided an interesting insight into how thymic development influenced the peptide recognition characteristics of the mature T cell repertoire. Differences in antigenic peptide recognition characteristics of T cells that developed in mice that display a normal diverse array of self peptides in the thymus and T cells that develop in engineered mice that display only a single type of self peptide (but are otherwise genetically identical) were determined. For a few mature T cells isolated from conventional mice and the engineered mice, the corresponding stimulatory peptides were identified. For each peptide, the amino acid at each residue was mutated to the nineteen other possibilities. For each variant peptide, the response of the T cell that was stimulated by the original peptide was re-assessed. If half the mutations at a particular residue abrogated a stimulator response, the residue was labelled a "hot spot". The greater the number of hot spots in a peptide, the more sensitive T cell recognition is to point mutations (i.e., it exhibits a higher degree of specificity). The data revealed that T cells that develop in conventional mice exhibited a significantly higher degree of specificity compared to those that developed in the engineered mice (Fig. 3).

Let us first explore the possible mechanistic origins of how thymic development shapes the T cell repertoire by carrying out a numerical experiment that mimics the experiments described above. Most of the TCR residues that make contacts with a peptide are in the so called Complementary Determining Region 3 (CDR3) region. This is the region in which TCRs are most different from each other. Consider a very simplified model in which the residues of a TCR that make contacts with peptides and corresponding residues of the peptide are represented as strings of sites (Fig. 4). In this representation, we are ignoring the conformations of the CDR loops and the peptide. As we are focused on short peptides and just the peptide contact residues of the TCR, this Occam's razor like approximation may not be so bad. Each of the TCR's peptide contact residues interacts with a single residue on the peptide. Generalizing this model to include multiple interactions between TCR and peptide residues does not change the qualitative results described below []. We can generate a repertoire of TCRs expressed by thymocytes by picking amino acids for each peptide contact residue in individual TCRs according to the frequencies of occurrence of amino acids in the human proteome. Similarly, we can generate an ensemble of peptides that are displayed in the thymus. The simplest estimate of the interaction free energy between a TCR (with a sequence of peptide contact residues, $\vec{t}$) and the resulting peptide-MHC molecules (with peptide sequence, $\vec{p}$ can be represented as:

$$E^t(\vec{t},\ \vec{p}) = E_c^t + \sum_{i=1}^{L} U\ (t_{i,}\ p_i) \qquad (1)$$

where E_c^t is the interaction free energy between the TCR and the MHC, and the function, U, describes the interaction between the amino acids at the L peptide contact residues of the TCR (t_i) and the corresponding peptide amino acids (p_i). For now, we will use a sign convention wherein more positive values of E^t correspond to stronger binding. The experiments carried out by Huseby and co-workers focused on T cells restricted by a single MHC allele, but still its interaction free energy with different

TCRs will vary, and so E_c^t should be picked from a distribution of values. If the value of E_c^t is too close to the positive selection threshold (and hence also the negative selection threshold as E_p and E_n are not that far apart), then the TCR is likely to be negatively selected because interactions with the peptide are likely to make the total interaction free energy exceed E_n. Conversely, if the value of E_c^t is too low, the TCR is unlikely to be positively selected. Therefore, an intermediate value of E_c^t is most likely to be realized in mature T cells. To carry out a numerical experiment, we have to choose the function, U. One possible choice is the Miyazawa-Jernigan potential that describes interactions between amino acids, and which has been used to study protein folding. This potential tends to overemphasize hydrophobic effects because of their importance for forming the core of folded proteins. Hydrophobic effects should also be important for forming the TCR-peptide-MHC interface. As we will soon see, the major mechanistic insights provided by theoretical analyses of our simple model of thymic development can be obtained without assuming any particular form of U.

Numerical interrogation of the effects of thymic development on the specificity of the TCR repertoire can now be carried out as follows. As described above, generate a million TCR sequences, each corresponding to a thymocyte, and a specific number of peptides that a thymocyte encounters in the thymus. Different thymocytes will encounter different numbers of peptides in the thymus, but the average number, denoted by N, is likely to be similar. There is no experimental evidence for this statement, and so the effects of fluctuations in N are worth investigating. The interaction free energies of each thymocyte's TCR with the encountered peptides are then estimated using Eq. 1. As noted earlier, experiments show that, in mice, the negative selection threshold, E_n, is sharply defined, while the positive selection threshold is less so, but still relatively sharp (Fig. 2b). Since we seek qualitative answers, in the numerical calculations, both thresholds are taken to be sharp. Based on experiments in mice described earlier, the gap between E_p and E_n can be taken to be roughly of the order of 1-2 k_BT.

For a T cell with a particular TCR to successfully mature into a peripheral T cell, none of the interaction free energies with the encountered peptides must exceed E_n, and at least one should exceed E_p. By carrying out numerical calculations with different numbers of encountered self peptides (N), TCR repertoires that develop when different numbers of self peptides are displayed in the thymus can be generated. To examine the dependence of the specificity of such TCR repertoires for recognition of pathogenic peptides, we can numerically generate a panel of pathogenic peptides by picking their residues according to the frequency of occurrence of amino acids in a model pathogen, for example, *Listeria Monocytogenes*. One can then carry out the numerical analog of the experiments by Huseby and co-workers. The interaction free energy of a mature TCR with several pathogenic peptides is determined. Based on the experimental finding [] that the threshold free energy of binding to peptide-MHC molecules that results in activation of mature peripheral T cells is similar to the value of E_n, a TCR is considered to recognize a pathogenic peptide if the corresponding interaction free energy exceeds E_n. For peptides recognized by a mature TCR, each of its amino acids is mutated to its nineteen possible mutants, and binding affinity is reassessed using Eq. 1. Peptide residues wherein half the mutations result in a binding free energy below E_n are termed hot spots (as in experiments). These numerical calculations are carried out with many realizations of thymocyte sequences, the peptides displayed in the thymus, and pathogenic peptides used to assess specificity. Thus, statistics are collected.

Fig 5 shows results of the numerical experiments described above for different values of N. The numerical results mirror the experimental finding in that mature T cells that develop in a thymus with a larger diversity of displayed self peptides exhibit a higher degree of specificity for recognition of pathogenic peptides. Given that this simple model recapitulates this key experimental observation, the important question now is to analyze the model further to seek mechanistic principles underlying this result and make experimentally testable predictions based on this knowledge.

Analysis of thymic selection effects in terms of extreme value statistics

In order to facilitate our analyses of the effects of thymic selection on T cell specificity, let us briefly digress to introduce a salient aspect of statistics. Consider N independent Gaussian variables {r_1, r_2, ….r_N}. Let P_N (x) be the probability that upon sampling the N variables, the maximum value is less than or equal to x; i.e.,

$$P_N\,(x) = \left[1 - \overline{\overline{P_1}}\,(x)\right]^N;\ \overline{\overline{P_1}}\,(x) = \int_x^\infty dr\, p_1\,(r) \qquad (2)$$

where p_1 (r) is the Gaussian probability distribution of each random variable. The probability distribution, $p_N\,(x)$, that the largest number lies between x and x+ dx is obtained from Eq. 2 by differentiation to equal:

$$p_N\,(x) = N\left[1 - \overline{\overline{P_1}}\,(x)\right]^{N-1}\left(-\frac{d\overline{\overline{P_1}}}{dx}\right) = N\,p_1(x)\left[1 - \overline{\overline{P_1}}\,(x)\right]^{N-1} \qquad (3)$$

Differentiating Eq 3 with respect to x, and setting the resulting expression to zero provides a way to obtain the most likely value of the largest number, x*. Specifically, x* must obey

$$p_1'\,(x^*) - p_1'\,(x^*)\,\overline{\overline{P_1}}\,(x^*) + N\,p_1^2\,(x^*) - p_1^2\,(x^*) = 0 \qquad (4)$$

If N is large, we expect x* to be a large number. Therefore, the values of p_1 (x*), $p_1'\,(x^*)$, and $\overline{\overline{P_1}}\,(x^*)$ will be sampled from the tails of probability distributions, and so will be small. Thus, to leading order, Eq. 4 reduces to

$$p_1'\,(x^*) + N\,p_1^2\,(x^*) = 0 \qquad (5)$$

We can solve Eq. 5 for x* by noting that p_1 (x) is the Gaussian probability distribution, $p_1\,(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{x^2}{2\sigma^2}}$. This obtains:

$$\ln x^* + \frac{x^{*2}}{2\sigma^2} = \ln N + \ln\frac{\sigma}{\sqrt{2\pi}} \qquad (6)$$

For large N, and hence x*, we obtain:

$$x^* = \pm\sqrt{2\sigma^2 \ln N} \qquad (7)$$

We will make use of this result in analyzing the effects of thymic selection.

We can also compute the probability, $P_N(x)$, for the probability that the largest of N random variables is less than or equal to x. Toward this end, note first that Eq. 4 implies that

$$\overline{\overline{P_1}}(x^*) = \frac{p_1'(x^*) + N\,p_1^2(x^*) - p_1^2(x^*)}{p_1'(x^*)} = \frac{1}{N} \qquad (8)$$

The second equality in Eq. 8 is obtained by using Eq. 5. Since $P_N(x) = \left[1 - \overline{\overline{P_1}}(x)\right]^N$, and for large N, in the vicinity of x*, $\overline{\overline{P_1}}(x)$ is small, we can rewrite this expression to read $P_N(x) = \exp\left[-N\,\overline{\overline{P_1}}(x)\right]$. Keeping only the leading order term in the expansion of $\overline{\overline{P_1}}(x)$ around its value at x*, and substituting the expression for $p_1(x^*)$ using Eq. 7, we obtain:

$$P_N(x) = \exp[-(1 - N(x - x^*)\,p_1(x^*)] = \exp\left[-\left(1 - \frac{x - x^*}{\sqrt{2\pi\sigma^2}}\right)\right] \sim \exp\left[-\exp\left[-\frac{x - x^*}{\sqrt{2\pi\sigma^2}}\right]\right] \qquad (9)$$

The last equality is obtained by recognizing that x is close to x*. Eq. 9 is called the Gumbell distribution. Finally, note that the above analyses will hold also for the statistics of minimum, rather than maximum, values with minor modifications, such as which sign in Eq. 7 is pertinent.

In light of this introduction to the statistics of extreme values, let us return to thymic selection. Recall that if a T cell has to successfully mature, it must bind at least one self peptide with an affinity that exceeds E_p, and none with an affinity that exceeds E_n. This is tantamount to saying that the maximum value of the interaction free energies of the T cell with the N self peptides encountered in the thymus must lie between E_p and E_n. Therefore, the probability that a T cell with a sequence of peptide contact residues, $\vec{t}$, will successfully mature, $P^{mat}(\vec{t})$, is:

$$P^{mat}(\vec{t}) = \int_{E_p}^{E_n} p_N(x, \vec{t})\,dx \qquad (10)$$

But, we also know that $P^{mat}(\vec{t})$ is the probability of not being negatively selected minus the probability of not being positively selected; i.e.,

$$P^{mat}(\vec{t}) = \left[1 - \overline{\overline{P_1}}(E_n, \vec{t})\right]^N - \left[1 - \overline{\overline{P_1}}(E_p, \vec{t})\right]^N \qquad (11)$$

Elementary calculus shows that Eqs 10 and 11 imply that $p_N(x, \vec{t})$ is the exact differential of $\left[1 - \overline{\overline{P_1}}(x, \vec{t})\right]^N$. Therefore, it is given by Eq. 3; i.e.,

$$p_N(x, \vec{t}) = N\left[1 - \overline{\overline{P_1}}(x, \vec{t})\right]^{N-1}\left(-\frac{d\overline{\overline{P_1}}}{dx}\right) = N\,p_1(x, \vec{t})\left[1 - \overline{\overline{P_1}}(x, \vec{t})\right]^{N-1} \qquad (12)$$

Notice that the probability that any T cell will bind very strongly to an arbitrary peptide very strongly is small. This is because a molecule is unlikely to bind to most other molecules avidly. So, $[1 - \overline{\overline{P_1}}\,(x, \vec{t})]$ has the shape shown in Fig. 6a. Since p_1 (x) is Gaussian, $p_N\,(x, \vec{t})$ has the shape shown in Fig. 6b. As N increases, the high power to which $[1 - \overline{\overline{P_1}}\,(x, \vec{t})]$ is raised in Eq. 11 makes $p_N\,(x, \vec{t})$ become more sharply peaked. For large enough N, this in turn implies that the condition for a T cell to successfully mature and become a peripheral cell is that the most likely value of the largest of the N interaction free energies of a particular T cell with N self peptides (x*) lies between E_p and E_n.

Eq. 1 describes the interaction free energy between a TCR and a peptide-MHC molecule for the simple model which when numerically simulated recapitulates experimental results regarding the dependence of TCR specificity on the number of self peptides displayed in the thymus. Using this formula, we see that the random number (r in our development of extreme value statistics) that we are sampling is:

$$r = E_c^t + \sum_{i=1}^{L} \langle U\,(t_i) \rangle + \sum_{i=1}^{L} \delta\, U\,(t_i) \qquad (13)$$

where $< U\,(t_i) >$ is the average value of the interaction free energy of the amino acid at the i^{th} peptide contact residue of the TCR under consideration with all other amino acids; $\delta\, U\,(t_i)$ is the fluctuation around this average that describe interactions with particular amino acids. If $\delta\, U\,(t_i)$ is distributed in a Gaussian fashion, we can use our previous result for x* (Eq. 7) to obtain the most likely value of the maximum interaction free energy of a TCR with N self peptides in the thymus (E*) to be:

$$E^* = E_c^t + \sum_{i=1}^{L} \langle U\,(t_i) \rangle + \sqrt{2\sigma^2 \ln N} \qquad (14)$$

where σ^2 is the sum of the variances of the interaction free energies of the individual amino acids, t_i, that make up the peptide contact residues of the TCR under consideration. Eq. 14 makes a prediction. As the number of types of self peptides encountered in the thymus (N) increases, the value of E* increases. So, to prevent thymocytes from being negatively selected, the TCR they express will have to be characterized by smaller values of $\sum_{i=1}^{L} \langle U\,(t_i) \rangle$ as N increases. That is, for large values of N, thymocytes that mature successfully in the thymus are predicted to express TCRs whose peptide contact residues are comprised of amino acids that do not bind other amino acids very strongly. They cannot bind to other amino acids too weakly either as that would prevent thymocytes with such TCRs from being positively selected. So, Eq. 14 predicts that, when a sufficiently large diversity of self peptides are expressed in the thymus, mature T cells are statistically likely to express TCRs with peptide contact residues that bind moderately to other amino acids. In contrast, when N is small, as in the engineered mouse that displays a single type of self peptide in the thymus, T cells with peptide contact residues that bind to other amino acids more strongly are more likely to successfully mature into peripheral T cells. The results obtained from our numerical experiments with the Miyazawa-Jernigan potential are consistent with these predictions (Fig. 7). These results hold if the interactions with MHC molecules is relatively narrowly distributed for T cells that successfully mature (see earlier).

Importantly, the predictions emerging from Eq. 14 hold true for any potential function, U. We anticipate that since TCR-pMHC interactions result in the formation of an interface, amino acids characterized by higher hydrophobicity are likely to interact more strongly with other amino acids. Thus, in the arguments that follow we will use hydrophobicity as a proxy for strongly interacting amino acid.

The biophysical reason underlying the result described by Eq. 14 is easy to understand. If selection of thymocytes is mediated by only one type of self peptide-MHC molecule, for a thymocyte to mature, its TCR must bind to this single molecule with an affinity that lies between E_p and E_n. Two common ways in which this can happen are shown in Fig. 8. One way is for the TCR to bind relatively strongly to the MHC, and weakly to the peptide. Another is to bind moderately to the MHC, but the binding affinity to the single type of peptide is dominated by one or two very strong contacts. Both of these two types of TCRs would be negatively selected if the corresponding thymocytes encountered a large diversity of self peptides in the thymus. For the first type of TCR, an encounter with any peptide with which its peptide contact residues interacted moderately with the peptide amino acids would likely push its affinity past the negative selection threshold. Similarly, additional moderate contacts with amino acids of other peptides would lead to negative selection for the second type of TCRs noted above. The point to recognize is that avoiding being negatively selected by a diversity of self peptides encountered in the thymus imposes a strong constraint that is most likely to be satisfied by thymocytes with TCRs that bind moderately to the MHC molecule that restricts it, and whose peptide contact residues are comprised of amino acids that are moderately hydrophobic.

The prediction that T cells that mature in a thymus that displays a greater diversity of self peptides are likely to be characterized by moderately hydrophobic peptide contact residues, while those that develop in a thymus with a small diversity of self peptides are likely to contain more hydrophobic peptide contact residues, explains the higher specificity exhibited by the former type of T cells. The activation threshold for mature T cells is similar to E_n. Therefore, for a T cell with moderately hydrophobic peptide contact residues in its TCR to recognize a pathogenic peptide, the amino acids of the peptide must be among the more strongly binding complements of the TCR peptide contact residues. Otherwise, it would be difficult to achieve an affinity of the order of E_n. Also, given that each peptide contact residue is likely to interact moderately with other amino acids, each contact with the peptide is likely to contribute a significant fraction of the total binding affinity; i.e., the threshold binding affinity for recognition is achieved by multiple moderate interactions acting collectively (Fig. 9).

Because the recognized peptide's amino acids contacted by the TCR are likely to be among the stronger binding complements of the TCR's peptide contact residues, if a peptide amino acid is mutated, it will likely be to an amino acid that binds less strongly to the corresponding TCR peptide contact residue. This will result in a decrease in binding affinity for the mutant peptide. Since each contact contributes a significant fraction of the overall affinity for T cells that mature in a thymus with a diversity of expressed self peptides, this effect is likely to result in a significant percentage reduction in the binding affinity. The affinity threshold that leads to recognition in the periphery is sharply defined (we will explore the origin of this observation in Chapter 7), and so a significant reduction in binding affinity is likely to lower it to

below that required for recognition. Taken together, the statements above argue that, for T cells that matured in a thymus with a diverse set of self peptides, most point mutations to a peptide recognized in the periphery are likely to abrogate recognition.

In contrast, for T cells that matured in a thymus with a small diversity of expressed self peptides, such as the engineered mouse that displays a single type of self peptide in the thymus, the TCR-pMHC affinity is likely to be dominated by one or two highly hydrophobic residues in the TCR peptide contact residues (Fig 9), or via strong interactions with MHC. So, individual point mutations to the peptide amino acids are less likely to abrogate recognition. These predictions originating from Eq. 14 are in agreement with the experimental findings of Huseby and co-workers described earlier. Calorimetric experiments that we shall not describe here suggest that in the few cases that were studied indeed T cells that matured in the normal mouse were characterized by a greater number of important contacts, compared to T cells that developed in mice with a single type of self peptide expressed in the thymus.

The model described above also enables us to understand why T cells can simultaneously exhibit high specificity by recognition being likely to be sensitive to point mutations in a recognized peptide, while simultaneously exhibiting some degree of cross-reactivity. This is because shuffling the peptide amino acid sequence in a way that maintains the same number of moderate interactions with the TCR may result in recognition.

The arguments described above make another important point. The mechanism underlying T cell recognition of specific peptide-MHC molecules is not described by Emil-Fisher's lock and key metaphor. Rather, recognition is the consequence of statistical matching of the patterns of the amino acids of TCR's peptide contact residues and the amino acids of the peptide. This is perhaps not surprising because T cells evolved to recognize peptides derived from a diverse world of peptides, including those that may not have evolved when an individual is born. To close this section, we note that the mechanistic picture that emerges by our consideration of highly simplified models is undoubtedly further embellished by many molecular details that were ignored. These details undoubtedly matter for individual TCRs and the pMHC molecules that they interact with. It is important, however, to test whether the essence of the proposed mechanism is statistically correct.

Experimental tests of the model predictions

The veracity of the model described above can only be established by directly testing its main prediction – viz., that the peptide contact residues of mature T cells that develop in mice with a diversity of self peptides in the thymus express TCR whose peptide contact residues are statistically more likely to be enriched in moderately hydrophobic amino acids. Furthermore, it needs to be established that, as per the model, this is a consequence of stringent negative selection constraints. Huseby and co-workers carried out a set of experiments in mouse models to explore these issues.

The investigations began by considering two TCRs, called YAe62 and B3K506, both of which are restricted by a common MHC class II molecule. In a particular mouse strain, B3K506-expressing thymocytes successfully mature, while those expressing YAe62 TCRs are eliminated. In some, but not all,

instances, the TCR β chain is dominant in determining interactions with peptide-MHC molecules. Therefore, transgenic mice with either the β chain of YAe62 or B3K506 were bred. The β chains randomly pair with a diversity of α chains in these mice. The following differences were noted between thymocytes of these TCRs:

1] The cells with YAe62 TCRs displayed higher levels of downstream signaling markers, suggesting that they were more self-reactive. However, the total number of cells expressing either YAe62 or B3K506 that matured in to single positive T cells were the same. These results suggest that greater numbers of YAe62 T cells are deleted by negative selection.
2] By breeding transgenic mice that do not express MHC molecules, and thus cannot carry out positive or negative selection, pre-selection thymocytes expressing either YAe62 or B3K506 TCR were isolated. The YAe62 thymocytes were found to be more reactive to self peptides in *in vitro* assays. These results suggest that TCRs with the YAe62 β chain paired with diverse α chains exhibit a high degree of self-reactivity.
3] The CDR3 regions of the β chains of the two TCRs are found to differ primarily at two residues that contact the sixth and seventh residues of peptides; these residues on the TCRs were thus labeled p6 and P7. Analyses of crystal structures revealed that these two residues are located at the center of the TCR-pMHC interface. The P6 and P7 amino acids for YAe62 are phenylalanine and tryptophan, while they are serines for B3K506. Swapping the amino acids at the P6 and P7 residues of the two TCRs reversed the order of their self-reactivity. As phenylalanine and tryptophan are more hydrophobic than serine, these results indicated that the hydrophobicity of the peptide contact residues of TCRs play a role in determining the strength of reactivity to self peptide-MHC molecules.
4] Analyses of crystal structures of 53 different human and mouse TCR complexed with peptide-MHC molecules, with different β chains of varying lengths, revealed that all use either the P6 or P7 residues to contact the peptide-MHC molecule, and in 43 cases both residues on the TCR are peptide contact residues. These results suggested that sequencing diverse TCRs that develop in mice, and relating the frequency of different types of amino acids at the P6 and P7 residues to their functional properties and fates during thymic development should be informative.
5] Large scale sequencing studies revealed that pre-selection TCRs with more hydrophobic amino acids at the P6 and P7 residues promote self-reactivity. A scale of hydrophobicity of this doublet of residues was found to correlate statistically with promoting, reducing, or being neutral with respect to self-reactivity for TCRs with various types of V_β chains.

These results set the stage for studying the effects of relative hydrophobicity of the peptide contact residues of TCRs on fate during thymic development. One way to study these effects is to examine the enrichment or depletion of the frequency of amino acids of different hydrophobicity in the mature T cell repertoire compared to the pre-selection repertoire (see point # 2 above to see how the latter class of TCRs can be isolated). Experimental data (Fig. 10) shows that, for both CD4 and CD8 T cells, only moderately hydrophobic amino acids are enriched in the peptide contact residues of mature T cells compared to the preselection repertoire. These results are concordant with the qualitative predictions made by Eq. 14 and our numerical experiments.

Bim is a molecule that plays an important role in mediating apoptosis (death) of thymocytes. Mice with the gene that encodes for Bim removed at both loci (so called Bim knock out, or Bim-/- mice) exhibit defects in negative selection. When the frequencies of amino acids in mature and preselection TCRs were studied in Bim-/- mice, strongly hydrophobic amino acids were statistically enriched in the mature T cell repertoire (Fig. 11). Taken together, the results shown in Figs. 10 and 11 demonstrate that negative selection promotes the deletion of thymocytes with TCRs whose peptide contact residues are comprised of strongly interacting, or highly hydrophobic, amino acids. These results suggest that the simple analyses of thymic selection that was developed in the previous section provides qualitatively accurate mechanistic insights into the statistical consequences of thymic selection for the mature T cell repertoire and its function.

Additional experimental results are indicative, but not confirmatory, of the notion that a greater diversity of self peptides expressed in the thymus suppresses the frequency of highly hydrophobic residues in the peptide contact residues of mature TCRs. A common strain of mice that is employed in laboratory studies is the C57BL6 strain (hereon labeled B6), which was used in the studies that led to the results shown in Figs. 10 and 11. A particular strain of mice, called Non-obese diabetic (NOD) mice, spontaneously develop diabetes, an autoimmune disease. Fig. 12 shows that mature CD4 T cells in NOD mice have a higher frequency of highly hydrophobic residues in the P6 and P7 residues compared to B6 mice. In contrast, this feature is not observed for CD8 T cells (Fig. 12). This result suggests that the feature observed for CD4 T cells is not because of any generic selection defect in NOD mice. The susceptibility of NOD mice to diabetes has been linked to their possessing a particular allele of the MHC II gene which expresses the MHC class II molecule called IA^g. These results suggest that the enrichment of highly hydrophobic residues in the P6 and P7 residues of the NOD TCR is due to the IA^g MHC class II molecule expressed in NOD mice. Huseby and co-workers generated two sets of mice to further explore this hypothesis. One set of mice called NOD.b, had all the background genes of NOD mice, but the MHC genes were those of B6 mice. Another set, called B6.$H2^g$ had all the background genes of the B6 strain, but the MHC genes were those of the NOD strain. For these strains of mice, the phenomenon observed in Fig. 12 was reversed. The B6.$H2^g$ mouse strain exhibited a higher frequency of highly hydrophobic residues in the mature CD4 T cell repertoire compared to B6 mice, and similar frequencies as the NOD mice (Fig. 13). Other studies have shown that IA^g binds peptides in a more unstable manner, suggesting that they may present a smaller diversity of self peptides in the thymus. Thus, the results shown in Figs 12 and 13 are in harmony with the simple theory discussed above; i.e., T cells that are selected against fewer self peptides are more likely to have strongly interacting amino acids at peptide contact residues.

The simple theory may also be consistent with certain findings in humans. People with certain MHC genes are more likely to control infection by pathogens that are highly mutable (e.g., HIV and HCV). Human MHC genes are referred to as Human Leukocyte Antigens (HLA). Two alleles of MHC class I genes, HLA-B57 and HLA-B27, are statistically overrepresented in cohorts of patients who can control HIV and HCV infections to maintain low levels of virus without treatment. Interestingly, individuals with these HLA alleles are also statistically more likely to acquire T cell-mediated autoimmune diseases, such as spondylitis and psoriasis. In the next chapter, we will see that the peptides presented by these HLA alleles are from regions of the HIV proteome that are highly vulnerable to mutations, and that this is the

principal reason for the fact that people with these HLA alleles are more likely to control HIV infections. However, the work described so far suggests that another factor may also play a role in this regard, and may also explain why individuals with these genes have a higher propensity for autoimmunity.

Machine learning approaches have been applied to experimental data on the equilibrium binding constants for diverse peptides binding to MHC molecules. For MHC class I molecules, the predictions made by the machine learning algorithms are fairly accurate because they bind peptides whose lengths are fairly tightly constrained. The largest possible set of human peptides can be constructed by taking the linear proteome, and including all peptides that overlap by one amino acid. Application of machine learning programs showed that HLA B-57 and HLA B-27 molecules bind to fewer human peptides, thus constructed. Given this result, Eq. 14 predicts that mature T cells with TCRs restricted by these HLA molecules will be more likely to have more strongly hydrophobic residues in their peptide contact residues. These T cells would be more likely to be cross-reactive to point mutants of the peptides that they recognize. Highly mutable pathogens can evolve mutations to evade T cell responses that target peptides derived from the pathogen's proteome. T cells in individuals with HLA-B57 and HLA-27 are more likely to be cross-reactive to point mutations of peptides that they recognize, and so would be more likely to inhibit mutational escape of highly mutable pathogens from the T cell response. This factor may contribute to their higher likelihood of being better at controlling infection with highly mutable pathogens, such as HIV and HCV. At the same time, T cells in these individuals would be more likely to exhibit autoimmune responses, since they were subjected to negative selection against a smaller diversity of self peptides compared to individuals with other MHC genes. An interesting issue ripe for future studies is why most people with HLA B-57 and HLA B-57 genes do not develop spondylitis and psoriasis, while people with these genes are overrepresented in patients with these autoimmune diseases.

Quorum sensing as a means to enable the T cell repertoire to be relatively self tolerant and pathogen sensitive

Mature T cells did not encounter the complete set of all possible self peptides presented by MHC molecules in the thymus. Since the peptides presented by MHC molecules are short, it is highly unlikely that by encountering some fraction of self peptides in the thymus, the T cells have learned any information about long range correlations in the host proteome. That is, a self peptide-MHC molecule that a particular mature T cell did not encounter in the thymus is just as likely to cause it to respond as a randomly chosen pathogenic peptide. Why then does this self reactivity of the T cell repertoire not cause autoimmune diseases in all individuals? Analysis of the consequences of thymic selection presented below suggests that an answer to this question might originate in collective effects.

In all our developments so far in this chapter, we have used a convention wherein higher binding free energies for TCR-peptide-MHC interactions correspond to larger positive numbers. In the physical sciences, often the opposite sign convention is used. Just to make sure that the reader is reassured that use of either sign convention results in the same qualitative insights, in this section we will use a

convention wherein stronger binding free energies correspond to larger negative numbers. The probability that a T cell with sequence, $\vec{t}$, is activated ($P\ (\vec{t})^{act}$) by a randomly chosen peptide is:

$$P\ (\vec{t})^{act} = \int_{-\infty}^{E_a} p_1\ (x,\ \vec{t})\ dx \qquad (15)$$

where E_a is the binding free energy threshold required for T cell activation for mature T cells, and the other symbols have been defined earlier. Earlier we found that the strongest interaction, $E^*\ (\vec{t})$, between self peptides encountered in the thymus by a thymocyte that successfully matured must lie between E_p and E_n. Since the magnitude of the difference between E_p and E_n is small, and the activation threshold, E_a, for mature T cells is very close to E_n, as an approximation, let us take $E^*\ (\vec{t})$ to be roughly equal to E_a in Eq. 15. For stronger interactions being negative, the definition of $\overline{\overline{P_1}}\ (x)$ (Eq. 2) becomes:

$$\overline{\overline{P_1}}\ (x) = \int_{-\infty}^{x} dr\ p_1\ (r) \qquad (16)$$

Therefore, with the approximations noted above,

$$P\ (\vec{t})^{act} = \int_{-\infty}^{E^*\ (\vec{t})} p_1\ (x,\ \vec{t})\ dx = \overline{\overline{P_1}}\ \left(E^*\ (\vec{t})\right) \qquad (17)$$

The following analog of Eq. 4 is obtained from these two equations and the condition for the location of the maximum of the extreme value of thymocyte interactions with self peptides (now denoted by $E^*\ (\vec{t})$):

$$p_1'\ \left(E^*\ (\vec{t})\right) - p_1'\ \left(E^*\ (\vec{t})\right)\ \overline{\overline{P_1}}\ \left(E^*\ (\vec{t})\right) - N\ p_1^2\ \left(E^*\ (\vec{t})\right) + p_1^2\ \left(E^*\ (\vec{t})\right) = 0 \qquad (18)$$

If the thymocyte with TCR, $\vec{t}$, encountered a large diversity of self peptides in the thymus, as is normal, then the value of $E^*\ (\vec{t})$ is going to be small, and so as in our previous developments, the probabilities in Eq. 18 will be drawn from the tails of the corresponding distributions. Therefore, we can again conclude that

$$\frac{p_1^2\left(E^*\ (\vec{t})\right)}{p_1'\left(E^*\ (\vec{t})\right)} \sim \frac{1}{N} \qquad (19)$$

We can always write Eq. 17 as:

$$P\ (\vec{t};\ E^*(\vec{t}))^{act} = \int_{-\infty}^{E^*\ (\vec{t})} \exp[\ln p_1\ (x,\ \vec{t})]\ dx \qquad (20)$$

Because E* is small, we can expand the logarithm in Eq. 19 around $E^*\ (\vec{t})$ and carry out the integral to obtain:

$$P\,(\vec{t};\ E^*(\vec{t}))^{act} = \frac{p_1^2\left(E^*(\vec{t})\right)}{p_1'\left(E^*(\vec{t})\right)} \qquad (21)$$

Comparing Eqs 19 and 21, we conclude that $P\,(\vec{t};\ E^*(\vec{t}))^{act} \sim \frac{1}{N}$. But, this estimate is not quite correct because we know that E_p and E_n are not so closely spaced that $E^*(\vec{t})$ ~ E_n ~ E^a. In fact, we expect $E^*(\vec{t})$ to be weaker than E^a because it will be weaker than E_n since it must lie between E_p and E_n. So, we expect that $P\,(\vec{t})^{act}$ will be smaller than 1/N, which was estimated using E^a equals $E^*(\vec{t})$. This is because if the strongest interaction of the TCR, $\vec{t}$, with a sampling of self peptides is less than E^a, the probability of its binding energy being equal to E^a for a randomly chosen peptide after maturation will be smaller than that calculated assuming that the strongest interaction with self peptides is equal to E^a. So, we can estimate $P\,(\vec{t})^{act} \sim \frac{1}{KN}$, where K > 1. Nonetheless, if a large number of self peptides are encountered in the thymus, the probability that a mature T cell is activated is small.

The above arguments show that, if a T cell interacts with fewer self peptides in the thymus, then it will be more reactive to both self and pathogen-derived peptides as a mature T cell. But, is there a difference between the probability of a randomly picked T cell's ability to be activated by a self or pathogenic peptide? On average, there should be a difference because any randomly picked T cell has interacted with some fraction of self peptides, and not been negatively selected by them, and so will not be activated by them as mature T cells. This argument suggests that the probability that a T cell is activated by a randomly chosen self peptide and a randomly chosen pathogenic peptide can be estimated as follows:

$$p^{act}\,(self) = p^{act}\left[1 - \frac{N}{M}\right] = p^{act}\,x \qquad (21)$$

$$p^{act}\,(pathogen) = p^{act} \qquad (22)$$

As our estimate for $p^{act}\left(\vec{t}\right)$ does not depend upon the particular TCR, Eqs 21 and 22 apply to any T cell, and so we dropped the reference to a particular TCR sequence.

For the T cell repertoire to be able to respond to all pathogens, if there are T clones of T cells (each with a different receptor), T p^{act} > 1. This ensures that the mean number of outcomes is greater than 1; i.e., at least one T cell in the repertoire is activated by a randomly chosen pathogenic peptide. Since p^{act} is small, T must be large for this to be true, and indeed the average human has 10^7 T cells with distinct receptors in a human. Similarly, we can estimate the probability that the T cell repertoire is completely able to avoid autoimmune responses to every self peptide-MHC molecule. This probability is given by $[(1 - p^{act}\,x)^T]^M \sim \exp[-MT\,p^{act}\,x]$, for small p^{act}. Even if every T cell encounters all, but one, self peptide during development (N = M-1), this probability of completely avoiding autoimmunity is given by $e^{-Tp^{act}}$, which is small if the repertoire is to be able to respond effectively to diverse pathogens (T p^{act} > 1). Our arguments based on rough estimates lead us to the conclusion that a T cell repertoire that is effective at combating diverse pathogens cannot avoid having T cells that are activated by some self peptides unless every thymocyte encounters every possible self peptide-MHC molecule during thymic

development. The latter condition is simply not true. So, what prevents the T cell repertoire from being rampantly autoreactive even though it has T cells that are autoreactive to some self peptides?

The answer to this question may be that collective effects between T cells determine whether an effective T cell response occurs. For example, let us postulate that a minimum number of T cells, t, must be activated for an effective response. The origin of such a phenomenon may be that after T cell activation, various growth factors and cytokines are required for the cells to proliferate. Regulatory T cells are known to suppress T cell responses, and they consume such cytokines. When a sufficient number of T cells are activated, they could secrete enough cytokines and growth factors to overcome such suppressive effects. Thus, a T cell response would be predicated by activation of a threshold number of T cells, t. This hypothesis is similar to the concept of quorum sensing that results in the response of bacterial populations to various environmental stimuli. If quorum sensing also underlies an effective T cell response, we can calculate the probabilities associated with the two kinds of errors that the T cell repertoire might make. Namely, an effective T cell response is mounted against a self peptide-MHC molecule or that an effective response is not mounted against a pathogenic peptide. The probability that t or more T cells are activated by a self peptide is given by:

$$P\ (E_1) = \sum_{k=t}^{T}\ {}_{k}^{T}C\ (p^{act}\ x)^k\ (1 -\ p^{act}x)^{T-k} \qquad (23)$$

where E_1 refers to the error made by mounting a response to a randomly chosen self peptide, and ${}_{k}^{T}C$ is the number of ways of choosing k T cells from a total of T in the repertoire. Similarly, we can calculate the probability that the T cell repertoire makes the second kind of error (E_2) referred to above; viz., that it does not mount an effective response to a pathogen derived peptide. This probability equals:

$$P\ (E_2) =\ \sum_{k=0}^{t-1}\sum_{k=1}^{t}\ {}_{k}^{T}C\ (p^{act})^k\ (1 -\ p^{act})^{T-k} \qquad (24)$$

Fig. 14a shows graphs of Eqs 23 and 24 as a function of t. if the threshold, t, is small, the probability of reactivity to self (error of type 1) is large, while the probability of not reacting to pathogen (error of type 2) is small. Note that the graph for P (E_1) decays to zero faster with t as the number of peptides, N, that a thymocyte encounters in the thymus increases. These graphs indicate that, for large enough values of N, there exists a value of t = t*, for which it is possible for the T cell repertoire to avoid autoimmune responses while being responsive to pathogens. The controlling parameter here is N/M; i.e., the average fraction of the total number of self peptide-MHC molecules that a typical thymocyte encounters during development. If N/M is sufficiently large, or x in Eq. 21 is sufficiently small, the concept of quorum sensing by mature T cells would explain why we are able to mount effective T cell responses to diverse pathogens, while being largely self-tolerant in spite of possessing some autoreactive T cells.

While graphing Eqs. 23 and 24 as in Fig. 14a is conceptually instructive, if parameters were known, these equations are not sufficient for calculation of the value of t*. The value of t* would need to ensure that there was no response to any of the self peptides displayed on APCs with which T cells may interact in a typical tissue. The value of t* would also have to ensure that a response was mounted to at least one of

the pathogen's peptides displayed on APCs upon infection. If the typical number of self peptides encountered by T cells in a tissue is S, the probability of not mounting autoimmune responses, (P (~S), is:

$$P\ (\sim S) = \ (1 - \ P\ (E_1))^S \qquad (25)$$

Upon infection, if the number of pathogenic peptides typically expressed on APCs is I, the probability that a response to a typical pathogen will be mounted, P (I), is:

$$P\ (I) = (1 - \ P\ (E_2)^I) \qquad (26)$$

Graphs of Eqs. 25 and 26 are shown in Fig. 14 b. If x is sufficiently small, depending upon the values of S and I, it is likely that a range of values of t exists that simultaneously results in high values of the probabilities in Eqs 25 and 26. Thus, quorum sensing may be a mechanism that allows the T cell repertoire to be largely self-tolerant, while maintaining the ability to mount effective responses to typical pathogens.

Several lines of experimental evidence now suggest that a quorum of T cells need to be activated for an effective response []. The clearest evidence has been provided by Rohr and co-workers. They studied CD8 T cells, which adhere to APCs such as Dendritic Cells (DCs) using a receptor called LFA-1 that binds to a protein called ICAM-1 that is expressed on DCs. Upon interactions of T cells with DCs bearing a stimulatory pMHC, it was found that T cells cluster together. Importantly, this clustering seems to be driven by the expression of ICAM-1 on activated T cells, which then allows them to adhere to each other and form clusters. CD80 and CD86 are molecules that are expressed on APCs that can bind to CD28 and CTLA-4 molecules expressed on T cells. Some naïve T cells express CD86. It was found that, upon activation, CD80 expression levels increased on T cells themselves. Thus, T cells were able to bind to CD28 on neighboring T cells in the cluster. CD28 signaling is known to promote the secretion of several cytokines, including IL2, which is an important growth factor mediating T cell proliferation. Experiments then showed that clustered T cells that interact mutually through CD28 and CD80 and CD86 expand in an IL2 dependent manner. The experiments also demonstrated that CD28 and IL2 mediated signaling was also found to lower the expression of apoptotic factors in T cells. Thus, T cells mutually regulated their expansion and survival by forming a quorum of clustered activated Tcells. Furthermore, the amount of IL2 produced scaled with the density of activated T cells that formed a quorum. Ultimately, population expansion stopped after some time due to interactions between CTLA-4 and CD80 and CD86 on clustered T cells. Turning off the immune response after some time is very important for preventing uncontrolled expansion and an overactive immune response that can cause harm. Interestingly, the inhibitory regulation of T cells due to CTLA-4 signaling is known to turn off the expansion of T cells that respond to cancer cells displaying peptides derived from mutated human proteins in cancer cells. One of the biggest advances in cancer treatment relies on blocking the action of CTLA-4 and other similar molecules using antibodies that bind to them. This allows T cells to continue to target cancer cells.

The study by Rohr and co-workers provides strong evidence that the theoretically predicted concept of quorum sensing described above plays a role in regulating T cell expansion in response to pathogenic peptides. Much more work that integrates the effects of regulatory T cells, activation of T cells by self peptides, etc is required to dissect the role of quorum sensing in potentially preventing autoimmunity.