Statistical Physics of T-Cell Development and Pathogen Specificity

Citation: Kosmrlj, Andrej, Mehran Kardar, and Arup K. Chakraborty. “Statistical Physics of T-Cell Development and Pathogen Specificity.” Annu. Rev. Condens. Matter Phys. 4, no. 1 (April 2013): 339–360.

As Published: http://dx.doi.org/10.1146/annurev-conmatphys-030212-184325

Publisher: Annual Reviews

Persistent URL: http://hdl.handle.net/1721.1/89815

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Statistical physics of T cell development and pathogen specificity

Andrej Košmrlj

Department of Physics, Harvard University, Cambridge, MA 02138

Mehran Kardar

Department of Physics, Massachusetts Institute of Technology,
Cambridge, MA 02139,
email: kardar@mit.edu

Arup K. Chakraborty

Departments of Chemical Engineering, Chemistry, and Biological Engineering,
Massachusetts Institute of Technology, Cambridge, MA 02139,
Ragon Institute of MGH, MIT, and Harvard, Boston, MA 02129,
email: arupc@mit.edu

Key Words immune system, HIV, extreme value distribution, Hamiltonian minimization

Abstract In addition to an innate immune system that battles pathogens in a non-specific fashion, higher organisms, such as humans, possess an adaptive immune system to combat diverse (and evolving) microbial pathogens. Remarkably, the adaptive immune system mounts pathogen-specific responses, which can be recalled upon re-infection with the same pathogen. It is difficult to see how the adaptive immune system can be preprogrammed to respond specifically to a vast and unknown set of pathogens. Although major advances have been made in understanding pertinent molecular and cellular phenomena, the precise principles that govern many aspects of an immune response are largely unknown. We discuss complementary approaches from statistical mechanics and cell biology that can shed light on how a key component of the adaptive immune system, T cells, develop to enable pathogen-specific responses against vast and diverse pathogens. The mechanistic understanding that emerges has implications for how host genetics may influence the development of T cells with differing responses to HIV infection.

CONTENTS

INTRODUCTION 2

SPECIFICITY & DEGENERACY OF THE T CELL REPERTOIRE 6

Thymic Selection Model 7

Abundance of Weak Interactions in TCRs Selected Against Many Self Peptides 10

Selection Against Many Self Peptides Leads to Pathogen-Specific T Cells 12

THYMIC SELECTION AS EXTREME VALUE PROBLEM 14

Nature of Foreign Peptides Recognized by T Cells 20

AN ASPECT OF THE ROLE OF HOST GENETICS IN CONTROL OF HIV THAT MAY BE RELATED TO THYMIC DEVELOPMENT 23

FUTURE PROSPECTS 26

DISCLOSURE STATEMENT 28

ACKNOWLEDGMENTS 28

Appendix: BASIC IMMUNOLOGY 29

The Two Arms of the Adaptive Immune System 29

The T Cell Recognition Process 30

1 INTRODUCTION

The immune system of an organism combats invading pathogens, thereby protecting the host from disease. Jawed vertebrates, such as humans, have an *adaptive immune system* that enables them to mount pathogen-specific immune responses [1]. The importance of this response for human health is highlighted by the opportunistic infections that afflict individuals with compromised adaptive immune systems [e.g., those who have progressed to AIDS after infection with the human immunodeficiency virus (HIV)]. Many other diseases (e.g., multiple sclerosis and type I diabetes) are consequences of the adaptive immune system failing to discriminate between markers of self and non-self. The suffering caused by autoimmune diseases, and the need to combat diverse infectious agents, has motivated a great deal of experimental research aimed at understanding how the adaptive immune system is regulated. These efforts have led to many notable discoveries [2–10], yet a deep understanding of the principles that govern the emergence of an immune or autoimmune response has proven elusive. This is highlighted by the inability to rationally design vaccines against many scourges on the planet (such as HIV).

An important barrier for the development of mechanistic principles that describe adaptive immunity is that the pertinent processes involve cooperative dy-

The first two paragraphs of the INTRODUCTION section (with small modifications) are reprinted with permission from the Annual Review of Physical Chemistry, Volume 61 (c) 2010, by Annual Reviews, http://www.annualreviews.org.

namic events. The many participating components must act collectively for an immune or autoimmune response to emerge. Moreover, these processes span a spectrum of time and length scales that range from interactions between molecules in cells to phenomena that affect the entire organism; feedback loops between processes on different spatiotemporal scales are also important. It is often hard to intuit underlying principles from experimental observations because of the complexity of these hierarchically organized collective processes. The importance of stochastic effects further confounds intuition.

Statistical mechanics provides a conceptual framework and tools (theoretical models and associated computations) that relate microscopic stochastic events to emergent complex behavior. When these insights are coupled closely to biological experiments, underlying physical and chemical mechanisms can be elucidated. In this review, we describe a project that brings together statistical mechanics and cell biology to uncover new concepts in immunology.

The adaptive immune system is not preprogrammed to respond to prescribed pathogens, yet it mounts pathogen-specic responses against diverse microbes, and establishes memory of past infections (the basis of vaccination). T lymphocytes (T cells) play an important role in coordinating adaptive immune responses. We explore how the developmental processes in an organ called the thymus shape the repertoire of T cells such that adaptive immunity exhibits both remarkable pathogen specificity and the ability to combat myriad pathogens. For the benefit of the uninitiated readers, the following bullet points present a minimal introduction to the relevant features of the immune system relevant for our study; the Appendix provides a slightly expanded description of basic immunology.

- *T cells* are a type of white blood cell that originate in the bone marrow, mature in the thymus (a gland in front of the heart and behind the sternum), and move to other parts of the body (via blood and lymph vessels) to fight off infections (pathogens).

- *Epitopes* are protein fragments (peptides) that constitute the molecular signatures of pathogens recognized by T cells. Both self and pathogen proteins are routinely chopped into pieces within a cell (by protease enzymes and the proteasome). Some fragments (typically 8-15 amino acids long [1]) can bind to an other host protein – major histocompatibility complex (MHC) – and these peptide-MHC (pMHC) complexes are displayed on the surface of a cell. Each human inherits several different types of MHC proteins from parents; the differences between MHC types between individuals are implicated in transplant rejection. T cells inspect pMHC complexes and initiate immune response if specific foreign peptides (epitopes) are encountered (see Fig. 1a).

- *T cell receptors (TCRs)* are proteins expressed on the surface of T cells which bind to pMHC complexes. The presence of an epitope, and subsequent T cell response, are determined by the strength of the binding. During synthesis in the bone marrow, immature T cells (*thymocytes*) acquire distinct TCR sequences through a gene shuffling process. Each such sequence can potentially bind strongly to a small number of complementary peptides, and a large repertoire of T cells is thus required to ensure proper coverage of the space of potential epitopes.

- *Thymic selection:* Following synthesis, thymocytes move to the thymus [7, 12–16]), where they interact with a variety of self-pMHC molecules (few thousands of different types [13]; see Fig. 2). These self-pMHC are derived from diverse parts of the host proteome and expressed on the surface of thymic epithelial cells as well as macrophages and dendritic cells. For a thymocyte to exit the thymus and become part of the host's repertoire of T cells, it must pass two tests: (a) It must not be *negatively selected*; i.e. its TCR must not bind to any self-pMHC molecule it encounters with a binding free energy that exceeds a threshold for negative selection. (b) It must bind at least one self-pMHC molecule with a binding free energy that exceeds another threshold for *positive selection*. It is thought that negative selection serves to delete dangerous T cells that may be activated by self pMHCs and cause autoimmune disease. The positive selection process ensures that TCRs of selected T cells do bind to a host's MHC (this is called *MHC restriction*).

The focus of this review is to understand how developmental processes in the thymus shape a T cell repertoire that exhibits both remarkable pathogen specificity, as well as the ability to combat myriad pathogens. This puzzle of specificity/degeneracy is described in Sec. 2, where we also present a model of the thymic selection process. Computational studies of this model characterize the properties of the selected T cell repertoire, which in turn elucidate the mechanism behind their specificity/degeneracy for pathogens. In Sec. 3, the above model of thymic selection is solved analytically by employing methods from statistical physics, such as extreme value distributions and Hamiltonian minimization. Genetic studies show that people with certain types of MHC are more likely to control HIV infections. We argue that these MHC types may affect thymic selection in a way that influences the statistical properties of the selected T cell repertoire, and this may provide one contributing factor (of many) for more efficient control of HIV infection (Sec. 4).

2 SPECIFICITY & DEGENERACY OF THE T CELL REPERTOIRE

TCR recognition of pathogen-derived pMHC molecules is both highly specific and degenerate (Fig. 1b). It is specific because if a TCR recognizes a pMHC molecule, most point mutations of the peptide's amino acids abrogate recognition [18, 19]. However, a given TCR can also recognize diverse peptides [9, 20–24]. This specificity-degeneracy conundrum is made vivid by dividing the world of peptides into classes, with the members of each class having sequences that are closely related. For example, peptides within a class could differ by just one-point mutation. A TCR can discriminate quite well between peptides within a class of closely related peptide sequences (as point mutants of the peptides it recognizes are not recognized with high probability). But, at the same time, a given TCR can recognize some other peptides in other classes, which have quite distinct sequences.

These diverse, specific/degenerate, (Fig. 1b) as well as a largely self-tolerant TCR repertoire is designed during T cell development in the thymus (Fig. 2; [7, 12–16])). Signaling events, gene transcription programs, and cell migration during T cell development in the thymus have been extensively studied [7, 12, 13, 15, 16, 25–30]. Experiments by Huseby et al. [18, 19] provide important clues about how interactions with self-pMHC complexes in the thymus shape the peptide binding properties of selected TCR amino acid sequences, such that mature T cells exhibit their special properties. These experiments contrasted T cells developed in conventional mice that display a diverse array of self-pMHC com-

Some parts of Sec. 2 are reprinted with permission from Ref. [17] (c) 2011, by Springer Science+Business Media.

plexes (few thousands of types) in the thymus to mice engineered to express only one type of peptide in their thymus. For T cells developing in conventional mice, recognition of antigenic pMHC was found to be sensitive to most point mutations of a recognized foreign peptide's amino acids. In contrast, T cells selected in mice with only one type of peptide in the thymus were much more peptide-degenerate, with some T cells being tolerant to several point mutations of recognized foreign peptide amino acids. The thymic selection model [17, 31–33] presented next explains these results, and also sheds light on the broader question of how the thymus designs diverse TCR sequences that mediate specific/degenerate pathogen recognition.

2.1 Thymic Selection Model

The key initiating event in T cell activation is the binding of a TCR to a pMHC complex. The binding interface of a TCR is composed of a more conserved region that is in contact with the MHC molecule, and a highly variable (CDR3) region that makes the majority of contacts with the peptide [1]. Accordingly, we divide the TCR/pMHC interaction free energy into two parts: a more conserved part represented by a continuous variable, and a part that explicitly depends on the variable TCR peptide contact residues and peptide sequences of amino acids. The former is given a value E_c, which may be varied to describe different TCRs and MHCs. The latter is obtained by aligning the TCR and pMHC amino acids that are treated explicitly and adding (in the simplest incarnation of the model) pairwise interactions between corresponding pairs (see, Fig. 3a). If the amino acid sequences of the TCR and the peptide are represented by strings $\vec{t} \equiv$

$(t_1, t_2, \cdots, t_N)$ and $\vec{s} \equiv (s_1, s_2, \cdots, s_N)$ respectively, the interaction free energy is

$$E_{\text{int}}(E_c, \vec{t}, \vec{s}) = E_c + \sum_{i=1}^{N} J(t_i, s_i). \tag{1}$$

The contribution from the ith amino acid of the TCR (t_i) and the corresponding peptide residue (s_i), is indicated by the matrix element $J(t_i, s_i)$. For numerical purposes we use the Miyazawa-Jernigan amino acid interaction matrix [34] developed in the context of protein folding, but as emphasized later, the qualitative results do not depend on the form of J. The length of the variable TCR-peptide region is taken to be $N \sim 5$. This is the typical number of peptide amino acids in contact with the TCR; the remaining peptide residues are important for binding to the MHC groove and/or are buried within the groove.

Earlier versions of such "string models" of TCR–pMHC interactions were used to study thymic selection [28, 29, 35], but they did not make an explicit treatment of amino acids (e.g., a formal string of numbers, bits, etc. was used). These studies provided estimates for certain properties of the selected TCR repertoire (for example, the number of selected TCRs activated by a foreign peptide or a foreign MHC – as in an organ transplant) that are consistent with experimental estimates. They also showed that negative selection in the thymus increases TCR specificity for foreign peptides, but did not suggest any mechanistic explanation. Other string models of TCR-pMHC interactions [36–38], with analogies to spin-glass models, were used also to study vaccination strategies for viral diseases and cancers (for a review see Ref. [39]).

To model thymic selection, we first construct a set $\mathcal{S} \equiv \{\vec{s}\}$ of M peptides of length N representing the self peptides encountered in the thymus. Each self peptide is generated as a sequence of N amino acids, each randomly and independently picked with frequencies corresponding to the human proteome [31, 40] (using the mouse proteome does not change the qualitative results [31]). We next generate many candidate sequences for the peptide contact residues of TCRs, $\vec{t}$, also randomly with the same amino acid frequencies. To mimic thymic selection, TCR sequences that bind to any of the M self-pMHC too strongly ($E_{\text{int}} < E_n$, with more negative free energies corresponding to stronger binding) are deleted (negative selection). However, a TCR must also bind sufficiently strongly ($E_{\text{int}} < E_p$) to at least one self-pMHC to receive survival signals and emerge from the thymus (positive selection). Recent experiments show that the difference between the thresholds for positive and negative selection is relatively small (a few $k_B T$ [27]). The threshold for negative selection (E_n) is quite sharp, while the threshold for positive selection (E_p) is soft [27, 41]. Replacing soft thresholds with perfectly sharp thresholds at E_n and E_p does not change the qualitative behavior of the selected T cell repertoire [17, 31].

To completely specify the interaction free energy between a TCR and pMHC, the value of E_c needs to be discussed. Selected TCRs are expected to bind moderately to MHCs, because binding too strongly to MHC (large $|E_c|$) would result in negative selection with any peptide, while too weak (small $|E_c|$) results in TCR not being positively selected. Each human can have up to 12 different MHC types. A TCR that binds strongly to more than one MHC type is likely to be eliminated during negative selection. Therefore, only TCRs binding to a particular MHC type are considered. This is consistent with the fact that there are no firm reports of a TCR restricted by more than one MHC type within a single human. Variations in E_c for selected TCRs are expected to be small. A rough estimate on the bounds is obtained from the condition that the average interaction free energy between TCR and pMHC for selected TCRs should be between the thresholds for positive and negative selection, i.e.

$$E_n < E_c + N\overline{J} < E_p, \tag{2}$$

where $\overline{J}$ is the average value of interaction between amino acids. The two bounds ($E_{c,\max} = E_p - N\overline{J}$ and $E_{c,\min} = E_n - N\overline{J}$) ensure that average interactions enable a TCR to survive both positive and negative selection. Since it is enough that a TCR sequence is positively selected by *any one* of many self peptides, and not negatively selected by *all* M self peptides, the precise bounds for E_c are different, but one expects that the range of E_c values is still small; viz., $E_{c,\max} - E_{c,\min} \propto E_p - E_n$. Note that TCRs whose interactions with MHCs are too weak are unlikely to be properly positioned on MHCs, and hence will be unable to interact with the peptide. Thus, one cannot tune E_c to very low values to escape negative selection. We thus assign to every TCR sequence a random value of E_c chosen uniformly from the interval $(E_{c,\min}, E_{c,\max})$, and then proceed with computing the consequences of the selection process.

2.2 Abundance of Weak Interactions in TCRs Selected Against Many Self Peptides

First, we summarize the results of computational analyses regarding how thymic selection shapes TCR sequences and TCR interactions with MHC. The peptide contact residues of TCR sequences selected against many self peptides in mouse and humans ($M \sim 10^3$ [13]) are statistically enriched with weakly interacting amino acids (Fig. 3b; [17, 31, 32]), and TCRs with weaker binding to MHC (within the allowed range) are more likely to get selected (Fig. 3c; [17]). This is because negative selection imposes a strong constraint. When selected against many self peptides, TCR sequences with peptide contact residues containing strongly interacting amino acids (e.g., hydrophobic amino acids or those with flexible side chains), or TCRs that bind strongly to MHC, are more likely to have strong binding with at least one encountered self-pMHC and thus be negatively selected. This qualitative result is independent of details of the interaction potential J, or the sharpness of the thresholds for positive and negative selection (see Sec. 3; [17, 31, 32]). Using different interaction potentials only changes the identities of the amino acids that interact weakly or strongly, or the criterion used to define interaction strength.

The conclusion that the peptide contact residues of selected TCRs are enriched in weakly interacting amino acids is supported by the analysis of available crystal structures of TCR-pMHCs [31]: amino acid frequencies of peptide contacting residues on TCRs in these crystal structures were determined and compared to amino acid frequencies in the human proteome (assumed to be the relevant frequencies for TCRs before thymic selection [31]). Measured amino acid frequencies in the TCRs' peptide contact residues were found to be smaller than in the human proteome for the most strongly (IVYWREL, Ref. [42]) interacting amino acids and larger for the weakly (QSNTAG, Ref. [42]) interacting amino acids.

2.3 Selection Against Many Self Peptides Leads to Pathogen-Specific T Cells

Does the selected T cell repertoire lead to specific recognition of a pathogenic peptide? To study the specificity of mature T cells in peptide recognition, selected TCR sequences were challenged with a collection of many randomly generated pathogenic peptides whose amino acid frequencies correspond to *Listeria monocytogenes* [31, 43], a pathogen that infects humans and is cleared by a T cell response. TCR recognition of pathogenic peptide occurs if TCR-pMHC binding is sufficiently strong ($E_{\text{int}} < E_r$), where the recognition threshold in mouse experiments is such that $E_r \sim E_n$ [44]. For each TCR that recognizes a particular pathogenic peptide sequence, the specificity of recognition was tested as follows: Each site on the peptide was mutated to all other 19 possibilities, and recognition of the mutated sequence by the original TCR was assessed. If more than half the mutations at a particular site abrogated recognition by the same TCR, the site was labeled an "important contact." For each TCR-pMHC pair for which recognition occurred, the number of important contacts was determined, and the resulting histogram is plotted in Fig. 4a. The higher the number of important contacts, the more specific is the TCR recognition of pathogenic peptide. Small numbers of important contacts correspond to *cross-reactive TCRs* that are able to recognize many pathogenic peptide mutants.

In agreement with experiments [18, 19], this model finds that TCRs selected against many different self peptides are very specific, while TCRs selected against only one self peptide are more cross-reactive (Fig. 4a). Based on the amino acid composition of selected TCRs, we can provide a mechanistic explanation for the specificity/degeneracy of pathogen recognition (Fig. 4b). Because TCR peptide contact residues are enriched with weakly interacting amino acids, they can interact sufficiently strongly for recognition to occur only with pathogenic peptides that are statistically enriched in amino acids that are the stronger binding complements of the peptide contact residues of the TCR (Fig. 3d). Such TCR-peptide pairs rely on many weak to moderate interactions which sum up to provide sufficient binding strength for recognition. Each interaction contributes a significant percentage of the total binding affinity. If there is a mutation to an amino acid of a recognized peptide, it is likely to weaken the interaction it participates in (as recognized peptides are statistically enriched in amino acids that interact strongly with the TCR's amino acids). Weakening an interaction that contributes a significant fraction of the binding free energy is likely to abrogate recognition because the recognition threshold is sharply defined [27].

In contrast, TCR sequences selected against only one type of self-peptide have a higher chance of containing strongly interacting amino acids (Fig. 3). Such TCRs can recognize a lot more pathogenic peptides including those composed of weakly or moderately interacting amino acids. In many cases mutating such amino acids on the peptide does not prevent recognition of the same TCR because a small number of strong contacts dominate recognition (Fig. 4a and experiments [19]). Unless these specific contacts are disrupted by mutations to the peptide, recognition is not abrogated. Accordingly, TCR recognition of pathogenic peptides is more cross-reactive. When selected against fewer types of self-peptides, TCRs that bind strongly to MHC can escape (Fig. 3c). Thus in this case the escape of TCRs that bind strongly or moderately to more than one MHC type (or MHC with mutations) might also be possible, leading to more cross-reactivity to MHC types (or substitutions of MHC amino acids [18]).

This mechanism for TCR-pMHC specificity is distinct from Fischer's [45] lock-and-key metaphor. Interactions between the TCR and the MHC dock the TCR over its ligand in essentially the same orientation [46, 47] – this may be analogous to shape complementarity, but it is not peptide specific. The complementary residues of the TCR then scan the peptide to assess if there is a sufficient number of moderate interactions to mediate recognition (Fig. 4b). An appropriate metaphor may be that the TCR peptide contact residues scan a bar code, and if there are a sufficient number of lines of moderate width (moderate TCR-peptide interactions), then recognition is posible. This statistical view of TCR specificity for pathogen may describe the initial step of binding, which may then allow modest conformational adjustments, leading to stronger binding [20]. This view is consistent with experiments suggesting a two-stage model for TCR-pMHC binding [48].

The statistical view of TCR-pMHC recognition also make degeneracy or cross-reactivity to peptides with different sequences the flip side of the coin. Although point mutations can abrogate recognition with high probability, making a number of changes to the peptide sequence such that a sufficient number of moderate interactions is still obtained will allow recognition by the same TCR (Fig. 4b). This may also be why two peptides with different sequences and conformations in the MHC groove can be recognized by the same TCR [20].

3 THYMIC SELECTION AS EXTREME VALUE PROBLEM

Interestingly, the thymic selection model presented in Sec. 2.1 can be solved *exactly* in the limit of long peptide sequences ($N \to \infty$) [17, 32]. A T cell expressing TCR with string $\vec{t}$ is selected in the thymus if its strongest interaction with a set $\mathcal{S}$ of M self-pMHCs is between the thresholds for negative and positive selection, i.e.

$$E_n < \min_{\vec{s}\in\mathcal{S}}\{E_{\text{int}}\left(E_c, \vec{t}, \vec{s}\right)\} < E_p. \tag{3}$$

Equation (3) casts thymic selection as an extreme value problem [49], enabling us to calculate the probability $P_{\text{sel}}(E_c, \vec{t})$ that a TCR sequence $\vec{t}$ is selected in

Parts of Sec. 3 are reprinted with permission from Ref. [32] (c) 2009, by American Physical Society, and from Ref. [17] (c) 2011, by Springer Science+Business Media.

the thymus. Let us indicate by $g(x|E_c,\vec{t})$ the probability density function (PDF) of the interaction free energy between the TCR $\vec{t}$ and a random peptide. The PDF $\Pi(x|E_c,\vec{t})$ of the strongest (minimum) of the M independent random free energies is then obtained by multiplying g with the probability of all remaining $(M-1)$ free energy values being larger, i.e.

$$\Pi\bigl(x|E_c,\vec{t}\bigr) = M\ g\bigl(x|E_c,\vec{t}\bigr)\left(1 - P\bigl(E < x|E_c,\vec{t}\bigr)\right)^{M-1}, \tag{4}$$

where $P(E\ <\ x|E_c,\vec{t})$ is the cumulative probability, and noting the multiplicity M of a particular interaction free energy being lowest. The probability that TCR $\vec{t}$ is selected is then obtained by integrating $\Pi(x|E_c,\vec{t})$ over the allowed range, as

$$P_{\text{sel}}\bigl(E_c,\vec{t}\bigr) = \int_{E_n}^{E_p} \Pi\bigl(x|E_c,\vec{t}\bigr)\,dx. \tag{5}$$

For $M \gg 1$, this extreme value distribution (EVD) converges to one of three possible forms, [49] depending on the tail of the PDF for each entry. Equation (1) indicates that in our case, as each interaction free energy is the sum of N contributions, $g(x|E_c,\vec{t})$ should be a Gaussian for large N, in which case the relevant EVD is the Gumbel distribution [49].

To obtain an explicit form for $\Pi(x|E_c,\vec{t})$, we model the set of self-peptides as M strings in which each amino acid is chosen independently. The probability f_a for selecting amino acid a at each site is taken to be the frequency of this amino acid in the self-proteome. For a specific TCR sequence $\vec{t}$, the average interaction free energy with self peptides then follows from Eq. (1) as

$$E_{\text{av}}(E_c,\vec{t}) = E_c + \sum_{i=1}^{N} \mathcal{E}(t_i), \tag{6}$$

with $\mathcal{E}(t_i) = \left[J(t_i,a)\right]_a$, where we have denoted the average over self amino acid frequencies by $\left[G(a)\right]_a \equiv \sum_{a=1}^{20} f_a G(a)$. Similarly, the variance of the interaction free energy is

$$V(E_c,\vec{t}) = \sum_{i=1}^{N} \mathcal{V}(t_i), \tag{7}$$

where $\mathcal{V}(t_i) = \left[J(t_i,a)^2\right]_a - \left[J(t_i,a)\right]_a^2$.

For very long peptide sequences (large N), we can approximate $g(x|E_c,\vec{t})$ with a Gaussian PDF with the above mean and variance. From standard results for the Gumbel distribution [49], we conclude that in the limit of $M \gg 1$, the peak of the distribution $\Pi(x|E_c,\vec{t})$ drifts to lower values as

$$E_0\left(E_c,\vec{t}\right) = E_{\text{av}}\left(E_c,\vec{t}\right) - \sqrt{2V\left(E_c,\vec{t}\right)\ln M}\,, \tag{8}$$

while its width is reduced to

$$\Sigma_0(E_c,\vec{t}) = \sqrt{\frac{\pi^2 V(E_c,\vec{t})}{12\ln M}}. \tag{9}$$

(Since the PDF $g(x|E_c,\vec{t})$ originates from a bounded set of free energies, it is strictly not Gaussian in the tails. Hence, once the extreme values begin to probe the tail of the distribution, the above results will no longer be valid. Indeed, in the limit when $M \sim \mathcal{O}(20^N)$, the EVD will approach a delta-function centered at the M–independent value corresponding to the optimal binding free energy.)

From Eq. (8) and the selection condition in Eq. (3) we see that as the number of self peptides, M, increases, the chance of negative selection does too. To counterbalance this pressure for large M, TCRs are enriched with weakly interacting amino acids in their peptide contact residues (small $\mathcal{E}(t_i)$ values), and with weaker interactions with MHC (small E_c value) (see Fig. 3). A similar effect relates to the variance of interactions (avoiding negative selection against many self peptides picks out TCRs with amino acid that exhibit a smaller variance in their interactions with other amino acids), but this tendency is less pronounced because of the square root. The preference for weak binding is independent of the potential J between contacting amino acids; different potentials merely reorder weak and strong amino acids.

Statistical mechanics suggests an analytic expression for the probability that a TCR sequence, $\vec{t}$, is selected according to Eq. (3) in the limit of large N and M. Remarkably the results seem to be accurate even for short peptides [17, 32]. A proper thermodynamic limit is obtained when $\{E_c, E_p, E_n\} \propto N$, and $\ln M \propto N$. The latter ensures that the peak of the EVD distribution, $E_0(E_c, \vec{t})$ in Eq. (8), is proportional to N. The same condition also implies that the width $\Sigma_0(E_c, \vec{t})$ in Eq. (9) is sharp and independent of N. The relation $\ln M = \alpha N$ can be justified from the expectation that M should grow proportionately to the proteome size P, while $N \propto \ln P$ to enable encoding the proteome. (The number of distinct peptide sequences of length N grows as 20^N, thus enabling encoding of proteomes with $P \leq 20^N$.) In this large N limit, the EVD is sufficiently narrow that the value of the optimal free energy can be precisely equated with the peak $E_0(E_c, \vec{t})$, and Eq. (3) for the selection condition can be replaced with

$$E_n < E_0\big(E_c, \vec{t}\big) < E_p. \tag{10}$$

The above thymic selection condition can now be interpreted as defining a micro-canonical ensemble of sequences $\vec{t}$, which are accepted if the value of the 'Hamiltonian' $E_0\big(E_c, \vec{t}\big)$ falls on the interval (E_n, E_p). In the large N limit, canonical and micro-canonical ensembles are equivalent and the probability is given by the Boltzmann weight of this Hamiltonian. More formally, the probability for TCR selection, $P_{\text{sel}}\big(E_c, \vec{t}\big)$, is obtained by using the least biased estimate, i.e. maximizing the entropy

$$S = \sum_{E_c, \vec{t}} P_{\text{sel}}\big(E_c, \vec{t}\big) \ln \big[P_{\text{sel}}\big(E_c, \vec{t}\big)\big] , \tag{11}$$

subject to a constraint that the average free energy

$$\left\langle E_0(E_c,\vec{t})\right\rangle = \sum_{E_c,\vec{t}} P_{\mathrm{sel}}\big(E_c,\vec{t}\big)\, E_0(E_c,\vec{t}), \tag{12}$$

is restricted to the interval (E_n, E_p). This leads to a probability for TCR selection governed by the Boltzmann-like weight [17, 32]

$$P_{\mathrm{sel}}\big(E_c,\vec{t}\big) \propto \left(\prod_{i=1}^{N} f_{t_i}\right) \rho(E_c) \exp\left[-\beta E_0(E_c,\vec{t})\right]. \tag{13}$$

Here $\{f_a\}$ and $\rho(E_c)$, are the prior frequencies of amino acids, and the distribution of E_c values before selection, whereas the effect of thymic selection is captured by the Boltzmann weight, with a Lagrange multiplier β determined by the condition $E_n < \left\langle E_0(E_c,\vec{t})\right\rangle < E_p$. Since the allowed values of $E_0(E_c,\vec{t})$ are bounded from above and below, the Lagrange multiplier β can be either negative or positive.

A difference compared to the standard micro-canonical ensemble is that the average free energy is constrained to an interval, rather than a precise value, necessitating a discussion on the choice of β. The possible values for $E_0(E_c,\vec{t})$ span a range from E_{min} to E_{max}, and the corresponding number of states form a bell-shaped curve between these extremes with a maximum at some E_{mid}. If $E_{\mathrm{mid}} > E_p$, to maximize entropy we must set $\left\langle E_0(E_c,\vec{t})\right\rangle = E_p$, and choose β accordingly. In this case, $\beta > 0$, positive selection is dominant and stronger amino acids are selected. If $E_{\mathrm{mid}} < E_n$, we must set β such that $\left\langle E_0(E_c,\vec{t})\right\rangle = E_n$, $\beta < 0$, negative selection is dominant and weaker amino acids are selected. For $E_n < E_{\mathrm{mid}} < E_p$, we must set $\beta = 0$ and there is no modification due to thymic selection.

Finally, we note that due to the appearance of $\sum_{i=1}^{N} \mathcal{V}(t_i)$ under the the square root term, Eq. (8) corresponds to an *interacting Hamiltonian* in which variables at different sites are apparently not independent. This is, however, not the case as the ‘interaction’ is easily removed by standard procedures such as Legendre transforms or Hamiltonian minimization [50], as follows: We need to solve a ‘Hamiltonian’ $\mathcal{H}(U,V)$ which depends on two extensive quantities $U=\sum_{i=1}^{N}\mathcal{E}(t_i)$ and $V=\sum_{i=1}^{N}\mathcal{V}(t_i)$. The corresponding partition function can be decomposed as $Z=\sum_{U,V}\Omega(U,V)e^{-\beta\mathcal{H}(U,V)}$, but can be approximated with its largest term. Note that the same density of states $\Omega(U,V)\equiv e^{S(U,V)/k_B}$ appears, irrespective of the specific form of $\mathcal{H}(U,V)$. In particular, the choice

$$\mathcal{H}_0(U,V)=E_c+U-\gamma V-\frac{\ln M}{2\gamma}=E_c+\sum_{i=1}^{N}[(\mathcal{E}(t_i)-\gamma\mathcal{V}(t_i)]-\frac{\ln M}{2\gamma}, \tag{14}$$

corresponds to a set of non-interacting variables, with

$$P_{\mathrm{sel}}\big(E_c,\vec{t}\big)\propto\rho(E_c)\exp[-\beta E_c]\prod_{i=1}^{N}\left\{f_{t_i}\exp\left[-\beta\left(\mathcal{E}(t_i)-\gamma\mathcal{V}(t_i)\right)\right]\right\}, \tag{15}$$

for which thermodynamic quantities (such as entropy) are easily computed. By judicious choice of γ we can then ensure that the same average free energy appears for $\mathcal{H}_0(E_c,\vec{t})$ and our $E_0(E_c,\vec{t})$. Using Legendre transforms, which is equivalent to minimizing $\mathcal{H}_0(E_c,\vec{t})$ with respect to γ, one finds that the required $E_0(E_c,\vec{t})$ is obtained by setting

$$\gamma(\beta)=\sqrt{\frac{\ln M}{2N\left\langle\mathcal{V}\right\rangle}}, \tag{16}$$

where $\langle\cdots\rangle$ refers to the average with the non-interacting weights in Eq. (15).

In practice we determine parameters β and γ as follows: Since the average free energy $\big\langle E_0(E_c,\vec{t})\big\rangle$ is a monotonic function of β, we use a bisection method to find the appropriate β that correspond to the specified value of the average. In order to do that, we need to discuss how to evaluate the average free energy for a particular β. First we use a bisection method to find a self-consistent value of γ from Eq. (16), and then calculate the average free energy using the Boltzmann weight in Eq. (15). We thus find β_p and β_n corresponding to $\big\langle E_0(E_c,\vec{t})\big\rangle=E_p$

and $\langle E_0(E_c, \vec{t}) \rangle = E_n$ respectively. Based on earlier discussion, we set $\beta = \beta_p$, when $0 < \beta_p < \beta_n$; $\beta = 0$, when $\beta_p < 0 < \beta_n$; and $\beta = \beta_n$, when $\beta_p < \beta_n < 0$.

Figure 5a depicts the variation of β as a function of $\ln(M)/N$ and the threshold for negative selection E_n, with $(E_p - E_n)/N = 0.5 k_B T$. With the thus obtained parameters β and γ we find the amino acid frequencies of selected TCRs as

$$f_a^{(\mathrm{sel})} = \frac{f_a \exp\left[-\beta(\mathcal{E}(a) - \gamma \mathcal{V}(a))\right]}{\sum_{b=1}^{20} f_b \exp\left[-\beta(\mathcal{E}(b) - \gamma \mathcal{V}(b))\right]}, \tag{17}$$

and the distribution of selected TCRs' interactions with MHCs as

$$\rho^{(\mathrm{sel})}(E_c) = \frac{\rho(E_c)\exp[-\beta E_c]}{\int_{E_{c,\min}}^{E_{c,\max}} \rho(E)\exp[-\beta E] dE}. \tag{18}$$

The above analytic expressions agree very well with numerical results from computer simulations of short peptides ($N = 5$) presented in the previous Sec. [17, 32].

3.1 Nature of Foreign Peptides Recognized by T Cells

After T cells complete thymic selection, a set $\mathcal{T}$ of TCRs, K in number, is available to respond to pathogens. A T cell recognizes infected cells when its TCR binds sufficiently strongly ($E_{\mathrm{int}} < E_n$) to foreign pMHC. This means that a foreign peptide of sequence $\vec{s}$ is recognized by some TCR if its strongest interaction with the set of TCRs exceeds the threshold for recognition, i.e.

$$\min_{\vec{t} \in \mathcal{T}} \left\{ E_{\mathrm{int}}\left(E_c, \vec{t}, \vec{s}\right)\right\} < E_n, \tag{19}$$

where the minimization is over the set of K TCRs (each with given E_c and $\vec{t}$) selected in the thymus.

Equation (19) casts recognition of foreign peptides as another extreme value problem. If we model the set $\mathcal{T}$ as K strings in which each amino acid is chosen independently with frequencies $f_a^{(\mathrm{sel})}$ (i.e. ignoring correlations among different positions on one string, and also between strings), then in the limit of large $K \gg 1$, the extreme value distribution is sharply peaked around [17]

$$E_0^*\left(\vec{s}\right) = \langle E_c \rangle + \sum_{i=1}^{N} \mathcal{E}^*(s_i) - \sqrt{\left(2\ln K\right)\left[\langle E_c^2 \rangle_c + \sum_{i=1}^{N} \mathcal{V}^*(s_i)\right]}, \tag{20}$$

and its width is

$$\Sigma_0^*(\vec{s}) = \sqrt{\frac{\pi^2\left[\langle E_c^2 \rangle_c + \sum_{i=1}^{N} \mathcal{V}^*(s_i)\right]}{12\ln K}}. \tag{21}$$

As $\ln K \propto N \to \infty$, the distribution becomes vary narrow and the condition for recognition of foreign peptides becomes

$$E_0^*\left(\vec{s}\right) < E_n. \tag{22}$$

The mean $\mathcal{E}^*(s_i)$ and the variance $\mathcal{V}^*(s_i)$ of the amino acid interaction free energies are obtained as in the previous section after replacing f_a with $f_a^{(\mathrm{sel})}$. The mean $\langle E_c \rangle$ and the variance $\langle E_c^2 \rangle_c = \langle E_c^2 \rangle - \langle E_c \rangle^2$ of selected TCR interactions with MHCs are obtained using $\langle X \rangle = \int_{E_{c,\min}}^{E_{c,\max}} X\rho^{(\mathrm{sel})}\left(E_c\right)\exp\left[-\beta E_c\right] dE_c$, with $\rho^{(\mathrm{sel})}\left(E_c\right)$ given in Eq. (18).

Repeating the reasoning of the previous section, the probability for a sequence $\vec{s}$ to be recognized is governed by the Boltzmann weight $P_{\mathrm{rec}}(\vec{s}) \propto \left(\prod_{i=1}^{N} \tilde{f}_{s_i}\right)\exp\left[-\beta^* E_0^*\left(\vec{s}\right)\right]$, where $\left\{\tilde{f}_a\right\}$ are prior frequencies of amino acids in the pathogen proteome, while the effect of TCR recognition is captured by the parameter β^*. As before, we introduce a new Hamiltonian $H_0^*\left(\vec{s}\right) = \langle E_c \rangle - \gamma^* \langle E_c^2 \rangle_c + \sum_{i=1}^{N}\left[\mathcal{E}^*(s_i) - \gamma^* \mathcal{V}^*(s_i)\right] - \ln K/\left(2\gamma^*\right)$, and to ensure the same average free energies, $\langle E_0^*\left(\vec{s}\right)\rangle = \langle H_0^*\left(\vec{s}\right)\rangle$, we set $\gamma^*\left(\beta^*\right) = \sqrt{\ln K/\left(2\langle E_c^2 \rangle_c + 2N\langle \mathcal{V}^* \rangle\right)}$. Finally, β^* is determined by constraining $\langle E_0^*\left(\vec{s}\right)\rangle < E_n$, while maximizing entropy. If $\beta^* > 0$, only foreign peptides with stronger amino acids are recognized. If $\beta^* = 0$, recognized peptides are not enriched or attenuated in strongly interacting amino acids. Note that unlike the parameter β for thymic selection of T cell receptors, β^* cannot be negative as there is no lower free energy bound for recognition in Eq. (22). The amino acid frequencies of recognized foreign peptides are then

$$\tilde{f}_a^{(\mathrm{rec})} = \frac{\tilde{f}_a \exp\left[-\beta^*(\mathcal{E}^*(a) - \gamma^*\mathcal{V}^*(a))\right]}{\sum_{b=1}^{20} \tilde{f}_b \exp\left[-\beta^*(\mathcal{E}^*(b) - \gamma^*\mathcal{V}^*(b))\right]}. \tag{23}$$

Figure 5b depicts variation of β^* as a function of the number of selected TCRs (K), the number of self peptides (M) against which TCRs were selected, and the threshold for negative selection E_n with $(E_p - E_n)/N = 0.5k_BT$. Notice, that in order for selected TCRs to recognize many foreign peptides (i.e. small value of β^*), we must have $K \gg M$ (i.e. a lot more selected TCRs than self-peptides presented in the thymus). This is consistent with biological values of $K \sim 10^9$ T cells [1] and $M \sim 10^3$ self peptides [13] in humans.

Equation (23) does not agree as well with the numerical results of computer simulations for short peptides ($N = 5$), as the corresponding ones for the selected TCR sequences presented before. The reason for the discrepancies is likely in the incorrect assumption that the selected TCR sequences are uncorrelated for small $N = 5$ [17]. However, the qualitative behavior of the parameter β^* as other model parameters are varied is expected to remain valid (Fig. 5b).

4 AN ASPECT OF THE ROLE OF HOST GENETICS IN CONTROL OF HIV THAT MAY BE RELATED TO THYMIC DEVELOPMENT

Each individual inherits a particular set of MHC molecules (up to six types of each MHC class I and class II protein) from their parents. Insights in to how

Some parts of Sec. 4 are reprinted with permission from Ref. [33] (c) 2010, by Nature Publishing Group.

thymic development shapes the T cell repertoire suggest a previously unknown aspect of how these differences in host genetics can influence the ability of humans to combat infectious diseases (such as HIV). HIV is a highly mutable and rapidly replicating virus that infects human T cells (among other cell types). HIV infection initially leads to acute high level viremia (the measurable presence of virus in the bloodstream), which is subsequently reduced to lower levels by the immune system. Without therapy, most patients experience a subsequent increase in viral load, and ultimately the development of AIDS. AIDS is associated with the occurrence of opportunistic infections because of the degradation of the immune system (T cells). Viremia levels and time to disease vary widely, and the differences correlate with the expression of different MHC class I molecules (as reviewed in Ref. [51]). Rare individuals ("elite controllers") maintain very low levels of HIV without therapy, thereby making disease progression and transmission unlikely. Certain MHC types appear more in elite controllers, with the highest association observed for the so-called HLA-B57 [52, 53]. While many complex factors may be at play, this fact suggests the involvement of T cells in viral control, since T cells activated by MHC bound viral peptides play an important role during various phases of disease [54–59]. T cells in people with different MHC genes could influence viral control in diverse ways. For example, it is known that MHC molecules associated with control present peptides derived from the HIV proteome that are vulnerable to mutations [51], especially because of collective effects of multiple simultaneous deleterious mutations [60]. Thus, T cells in people with these MHC molecules are thought to target more vulnerable regions of HIV, thereby hindering mutational escape from the host immune pressure.

A puzzling finding is that the MHC molecules most associated with enhanced control of HIV, HLA-B57 and HLA-B27, are also associated with increased proclivity for certain autoimmune disorders. Indeed, HLA-B57 has been associated with autoimmune psoriasis [61] and hypersensitivity reactions [62], and HLA-B27 with ankylosing spondilytis [63]. The understanding of the role of thymic development in shaping the T cell repertoire that has emerged from experimental and theoretical studies (vide supra) may explain the mechanistic origins of these observations.

Bioinformatics algorithms [64] based on experimental data predict whether a particular peptide will bind to a given MHC molecule [33]. Using these algorithms, the fraction of peptides derived from the human proteome [65] that bind to various MHC molecules were computed. Of the roughly 10^7 unique peptide sequences, only 70,000 are predicted to bind to HLA-B57, while 130,000 bind to a typical HLA-B molecule, and 180,000 bind to HLA-B7 (an MHC type that is associated with faster progression to AIDS) [33].

The intrinsic differences in self-peptide binding among MHC molecules can be important during development of immature T cells in the thymus. As fewer self peptides are able to bind to HLA-B57 molecules, a smaller diversity of self pMHC are encountered by HLA-B57–restricted T cells in the thymus. Thus, as described earlier, HLA-B57–restricted T cells are likely to be more cross-reactive to point mutants of targeted viral peptides than T cells restricted by MHC types that present a greater diversity of self peptides (Fig. 4a). This finding is supported by experiments measuring the cross-reactivity of T cells from people with diverse MHCs for HIV peptides [66–68].

A model of host–HIV dynamics showed that a repertoire of T cells more cross-reactive to point mutants of targeted epitopes results in better control of HIV infection [33]. This is because such T cells can exert immune pressure on the infecting strain and mutants that rapidly emerge to escape the immune pressure more effectively. Thus, it was predicted that HIV-infected individuals with MHC types that bind fewer self peptides are more likely to control viral loads to low values. Supporting these predictions, in a large cohort of HLA-typed individuals, experiments showed that the relative ability of HLA-B MHC types to control HIV infection correlates with their peptide-binding characteristics that affect thymic development (Fig. 6; [33]). Furthermore, there is also evidence that the immune response in individuals with the HLA-B27 gene that control HIV exhibits greater proportion of cross-reactive T cells than HLA-B27 positive individuals who do not control HIV [69]. Even though we do not fully understand why individuals with the HLA-B27 gene exhibit different proportions of activated cross-reactive T cells upon HIV infection, its effects on the control of HIV support our conclusions. Undoubtedly, many complex factors influence the relationship between MHC type and disease outcome. The effect of the factor related to differential thymic selection should be greatest for MHC molecules that bind relatively few (for example, HLA-B57) or many (for example, HLA-B7, -B35, -B8) self peptides.

Superior control of viral load due to the greater precursor frequency and cross-reactivity of T-cell repertoires restricted by MHC molecules that bind to few self peptides (for example, HLA-B57), should also confer protection against diseases caused by other fast-mutating viruses. Indeed, HLA-B57 is protective against hepatitis C virus, HCV [70], another highly mutable viral disease in which T cells are important. Also, HLA-B8, which binds a greater diversity of self peptides [33], is associated with faster disease progression in HCV [71] and HIV [66]. Thus, the correlation between the diversity of peptides presented in the thymus during T-cell development and control or progression of disease may be general.

The results we summarize above also point to a mechanistic explanation for the previously unexplained associations between HLA alleles that confer protection against HIV and autoimmune diseases. T cells restricted by MHC types that bind to few self peptides are subject to less stringent negative selection in the thymus, and should therefore be more prone to recognizing self peptides. This may explain the enhanced proclivity for autoimmune disorders in people with MHC genes that are also associated with superior control of HIV infections.

5 FUTURE PROSPECTS

In this review we address how simple statistical mechanical models can be used to shed light on certain aspects of the immune response. However, due to the highly complex characteristics of the adaptive immune system, many basic questions remain unresolved. The richness and intricacy of the problem invites a multitude of approaches from the physical and life sciences to uncover new principles. Below we discuss some additional questions pertinent to development and actions of the T cell repertoire where models similar to the ones presented here could lead to new insights.

Thymic selection attempts to remove dangerous T cells that could cause autoimmune disease. But thymic selection is not perfect and some autoreactive T cells may escape, possibly due to the fact that all possible self-peptide types are not expressed in the thymus, and that immature T cells spend a finite time in the thymus. The immune system thus has other protective mechanisms, e.g. certain 'regulatory T cells' suppress immune responses directed against self tissues. Are T cells that do not encounter certain self peptides in the thymus just as reactive to these peptides as those derived from pathogens? If so, how is autoimmunity suppressed, but not reactivity to pathogens? Most times, the immune system is very efficient at preventing autoimmune diseases, but what leads to a higher frequency of escape of autoreactive T cells which target cells of the nervous system and the pancreas in the case of autoimmune diseases like multiple sclerosis and type I diabetes? Why do the escaping autoreactive T cells attack only particular tissues? By adapting the thymic selection model to include the variability in expression levels of different types of self peptides in the thymus, one could potentially get insights into the last two questions. The escape probability of autoreactive T cells from the thymus can also be studied as diffusion in a random field of immobile traps [72].

Autoimmune diseases are correlated to a combination of genetic and environmental factors. People who express certain genes have a higher propensity for certain autoimmune diseases, but not everyone with these genes develops disease; e.g., most people with the inflammatory disease ankylosing spondylitis express HLA-B27 (a type of MHC), but most people expressing HLA-B27 do not develop ankylosing spondylitis. Certain genes and viral infections are known to increase the risk of triggering multiple sclerosis. Insights into these puzzles could emerge from stochastic dynamical models of host-pathogen interactions coupled with models of T cell development.

6 DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

7 ACKNOWLEDGMENTS

Our work in this area has greatly benefited from fruitful discussions and collaborations with Eric S. Huseby, Eugene I. Shakhnovich, Bruce D. Walker, and Herman N. Eisen. This work was supported by the Ragon Institute (A.K.C., A.K.), National Institutes of Health (NIH) Grant No. 1-PO1-AI071195-01 (A.K.C., M.K.), NSF Grant No. DMR-12-06323 (M.K.), and a NIH Directors Pioneer award (to A.K.C.).

A Appendix: BASIC IMMUNOLOGY

Higher organisms are constantly exposed to infectious microbial pathogens, yet rarely develop disease. This is because the variety of cells that comprise the innate immune system are efficient in controlling pathogenic microorganisms. The components of the innate immune system respond to common features of diverse microorganisms, but are not specific for individual pathogens. Some bacteria and many viruses can evade or overcome the innate mechanisms of host defense. The adaptive immune system mounts pathogen-specific responses against such invading microorganisms. Adaptive immunity also establishes memory of past infections, thereby conferring the ability to mount rapid immune responses to pathogens encountered previously. This immunological memory is the basis for vaccination.

A.1 The Two Arms of the Adaptive Immune System

The adaptive immune system has two arms, called cellular and humoral immunity. T lymphocytes (T cells) and B lymphocytes (B cells) are the key regulators of cellular and humoral immunity, respectively. T cells and B cells express immunoglobulin proteins on their surfaces, which are called T cell receptors (TCRs) and B cell receptors (BCRs), respectively. The genes encoding these receptors are inherited as gene segments that stochastically recombine during the synthesis of T cells and B cells in the bone marrow. Each gene assembled in a given lymphocyte is thus likely to be distinct, enabling the generation of a great diversity of T cells and B cells expressing different receptors. Different lymphocytes can

Parts of this Appendix are reprinted with permission from the Annual Review of Physical Chemistry, Volume 61 (c) 2010, by Annual Reviews, http://www.annualreviews.org.

potentially respond to specific pathogens as distinct receptors can potentially recognize (i.e., bind sufficiently strongly to) molecular signatures of specific foreign invaders. While the adaptive immune system can mount pathogen-specific responses against varied microbes, the number of possible pathogens suggests that one T cell (or B cell) for every possible pathogen is an unlikely scenario.

The diverse lymphocytes bearing different TCRs and BCRs generated in the bone marrow do not all become part of an organism's army of T cells and B cells. Rather, T cells and B cells undergo development processes that allow only a small fraction of the generated cells to become part of an organism's repertoire of lymphocytes. T cells develop the thymus (the T stands for thymus); B cells develop in the bone marrow (the B stands for bone marrow) and also, upon activation, in lymphoid organs.

A.2 The T Cell Recognition Process

Cells of the innate immune system (e.g., dendritic cells, macrophages) engulf pathogens (also called antigens) present in different parts of an organism's body. These cells are called antigen-presenting cells (APCs) because they express molecular signatures of the ingested antigens on their surface. Extracellular fluid from tissues, which contains pathogens or APCs harboring pathogens, drains into lymphoid organs (e.g., lymph nodes, spleen) via the lymphatic vessels. In lymphoid organs, lymphocytes can interact with pathogen-bearing APCs and pathogens and recognize them as foreign.

If a lymphocyte recognizes pathogens in a lymph node, a series of intracellular biochemical reactions occurs (called signaling) that results in gene transcription programs that cause the lymphocyte to become activated; i.e., it begins to proliferate and acquire the ability to carry out functions that can mediate an immune response. Activated lymphocytes thus generated, bearing receptors specific for the infecting pathogen, then leave the lymph node and enter the blood via lymphatic vessels. The lymphatic vessels enable lymphocytes to circulate among the blood, lymphoid organs, and tissues. When activated lymphocytes encounter the same pathogen's molecular markers in the blood or tissues (see below), they can carry out effector functions to eliminate the infection.

The BCRs and TCRs expressed on B cells and T cells can bind to molecular markers called *ligands*. B cells protect against pathogens in blood or *extracellular* spaces. The ligands of the BCR include proteins, fragments of proteins, and molecules on the surface of viruses or bacteria. T cells evolved to combat *intracellular* pathogens. Proteins synthesized by intracellular pathogens are cut up into short peptide fragments (typically 8-15 amino acids long [1]) by enzymes in cells harboring the pathogen. These peptide fragments may bind to the host's "major histocompatibility complex (MHC)" proteins. There are two kinds of MHC proteins, called class I and class II. Typically, a human will have up to six types of MHC class I proteins, and up to six types of MHC class II proteins. Pathogen-derived peptides (p) bound to MHC proteins are ultimately expressed on the surface of APCs (encountered by T cells in lymph nodes prior to activation) and infected cells (encountered by T cells patrolling blood and tisues); see Fig. 1a.

T cell recognition of a particular pathogen-derived pMHC implies that its TCR binds to it sufficiently strongly, leading to productive intracellular signaling and activation. The T cell signaling network does not respond progressively to increasing the stimulus (e.g., TCR-pMHC binding strength); rather, it only responds strongly above a threshold stimulus level [27, 41]. In the lymph nodes, T cells activated by peptides presented by MHC class II proteins proliferate and differentiate into many cell types called T helper cells, as they help activate B cells and perform other important functions. T cells activated by peptides presented by MHC class I molecules are called cytotoxic T lymphocytes (CTLs). When activated CTLs encounter cells in tissues that express the pMHC molecules that originally activated them, they can kill these cells by secreting various chemicals.

References

1. C. Janeway, P. Travers, M. Walport, and M. Shlomchik, Immunobiology (Garland Science, New York, 2004), 6th ed.
2. J. P. Allison, B. W. McIntyre, and D. Bloch, J. Immunol. **129**, 2293 (1982).
3. A. C. Chan, M. Iwashima, C. W. Turck, and A. Weiss, Cell **71**, 649 (1992).
4. D. P. Dialynas, D. B. Wilde, P. Marrack, A. Pierres, K. A. Wall, W. Havran, G. Otten, M. R. Loken, M. Pierres, J. Kappler, et al., Immunol. Rev. **74**, 29 (1983).
5. A. M. Gallegos and M. J. Bevan, Immunol. Rev. **209**, 290 (2006).
6. S. M. Hedrick, D. I. Cohen, E. A. Nielsen, and M. M. Davis, Nature **308**, 149 (1984).
7. K. A. Hogquist, T. A. Baldwin, and S. C. Jameson, Nat. Rev. Immunol. **5**, 772 (2005).
8. A. B. Irving and A. Weiss, Cell **64**, 891 (1991).
9. E. R. Unanue, Annu. Rev. Immunol. **2**, 395 (1984).

10. Y. Yanagi, Y. Yoshikai, K. Leggett, S. P. Clark, I. Aleksander, and T. W. Mak, Nature **308**, 145 (1984).

11. A. K. Chakraborty and A. Košmrlj, Annu. Rev. Phys. Chem. **61**, 283 (2010).

12. S. C. Jameson, K. A. Hogquist, and M. J. Bevan, Annu. Rev. Immunol **13**, 93 (1995).

13. O. M. Siggs, L. E. Makaroff, and A. Liston, Curr. Opin. Immunol. **18**, 175 (2006).

14. T. K. Starr, S. C. Jameson, and K. A. Hogquist, Annu. Rev. Immunol **21**, 139 (2003).

15. H. von Boehmer, I. Aifantis, F. Gounari, O. Azogui, L. Haughn, I. Apostolou, E. Jaeckel, F. Grassi, and L. Klein, Immunol. Rev. **191**, 62 (2003).

16. G. Werlen, B. Hausmann, D. Naeher, and E. Palmer, Science **299**, 1859 (2003).

17. A. Košmrlj, M. Kardar, and A. K. Chakraborty, J. Stat. Phys. pp. doi 10.1007/s10955–011–0403–8 (2011).

18. E. S. Huseby, J. White, F. Crawford, T. Vass, D. Becker, C. Pinilla, P. Marrack, and J. W. Kappler, Cell **122**, 247 (2005).

19. E. S. Huseby, F. Crawford, J. White, P. Marrack, and J. W. Kappler, Nat. Immunol. **7**, 1191 (2006).

20. H. N. Eisen and A. K. Chakraborty, Proc. Natl. Acad. Sci. USA **107**, 22373 (2010).

21. B. Hemmer, M. Vergelli, B. Gran, N. Ling, P. Conlon, C. Pinilla, R. Houghten, H. F. McFarland, and R. Martin, J. Immunol **160**, 3631 (1998).

22. G. J. Kersh and P. M. Allen, Nature **380**, 495 (1996).

23. I. S. Misko, S. M. Cross, R. Khanna, S. L. Elliott, C. Schmidt, S. J. Pye, and S. L. Silins, Proc. Natl. Acad. Sci. USA **96**, 2279 (1999).

24. J. Sloan-Lancaster and P. M. Allen, Annu. Rev. Immunol. **14**, 1 (1996).

25. J. A. M. Borghans, A. J. Noest, and R. J. de Boer, Eur. J. Immunol. **33**, 3353 (2003).

26. P. Bousso, N. R. Bhakta, R. S. Lewis, and E. Robey, Science **296**, 1876 (2002).

27. M. A. Daniels, E. Teixeiro, J. Gill, B. Hausmann, D. Roubaty, K. Holmberg, G. Werlen, G. A. Holländer, N. R. J. Gascoigne, and E. Palmer, Nature **444**, 724 (2006).

28. V. Detours, R. Mehr, and A. S. Perelson, J. Theor. Biol. **200**, 389 (1999).

29. V. Detours and A. S. Perelson, Proc. Natl. Acad. Sci. USA **96**, 5153 (1999).

30. A. Scherer, A. Noest, and R. J. de Boer, Proc. R. Soc. London Ser. B **271**, 609 (2004).

31. A. Košmrlj, A. K. Jha, E. S. Huseby, M. Kardar, and A. K. Chakraborty, Proc. Natl. Acad. Sci. USA **105**, 16671 (2008).

32. A. Košmrlj, A. K. Chakraborty, M. Kardar, and E. I. Shakhnovich, Phys. Rev. Lett. **103**, 068103 (2009).

33. A. Košmrlj, E. L. Read, Y. Qi, T. M. Allen, M. Altfeld, S. G. Deeks, F. Pereyra, M. Carrington, B. D. Walker, and A. K. Chakraborty, Nature **465**, 350 (2010).

34. S. Miyazawa and R. L. Jernigan, J. Mol, Biol. **256**, 623 (1996).

35. D. L. Chao, M. P. Davenport, S. Forrest, and A. S. Perelson, Eur. J. Immunol. **35**, 3452 (2005).

36. J. M. Park and M. W. Deem, Phys. A, Stat. Mech. Appl. **341**, 455 (2004).

37. M. Yang, J. M. Park, and M. W. Deem, Lect. Notes Phys. **704**, 541 (2006).

38. H. Zhou and M. W. Deem, Vaccine **24**, 2451 (2006).

39. M. W. Deem and P. Hejazi, Annu. Rev. Chem. Biomol. Eng. **1**, 247 (2010).

40. P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, et al., Nucleic Acids Res. **36**, D707 (2008).

41. A. Prasad, J. Zikherman, J. Das, J. P. Roose, A. Weiss, and A. K. Chakraborty, Proc. Natl. Acad. Sci. USA **106**, 528 (2009).

42. K. B. Zeldovich, I. N. Berezovsky, and E. I. Shakhnovich, PLoS Comput. Biol. **3**, 62 (2007).

43. I. Moszer, P. Glaser, and A. Danchin, Microbiol. UK **141**, 261 (1995).

44. D. Naeher, M. A. Daniels, B. Hausmann, P. Guillaume, I. Luescher, and E. Palmer, J. Exp. Med. **204**, 2553 (2007).

45. E. Fischer, Ber. Dtsch. Chem. Ges. **27**, 2985 (1894).

46. D. N. Garboczi, P. Ghosh, U. Utz, Q. R. Fan, W. E. Biddison, and D. C. Wiley, Nature **384**, 134 (1996).

47. K. C. Garcia, M. Degano, R. L. Staneld, A. Brunmark, M. R. Jackson, P. A. Peterson, L. Teyton, and I. A. Wilson, Science **274**, 209 (1996).

48. L. C. Wu, D. S. Tuot, D. S. Lyons, K. C. Garcia, and M. M. Davis, Nature **418**, 552 (2002).

49. M. R. Leadbetter, G. Lindgren, and H. Rootzen, Extremes and related properties of random sequences and processes. (Springer-Verlag, 1983).

50. M. Kardar, Phys. Rev. Lett. **51**, 523 (1983).

51. S. G. Deeks and B. D. Walker, Immunity **27**, 406 (2007).

52. S. A. Migueles, M. S. Sabbaghian, W. L. Shupert, M. P. Bettinotti, F. M. Marincola, L. Martino, C. W. Hallahan, S. M. Selig, D. Schwartz, J. Sullivan, et al., Proc. Natl. Acad. Sci. USA **97**, 2709 (2000).

53. T. Miura, C. J. Brumme, M. A. Brockman, Z. L. Brumme, F. Pereyra, B. L. Block, A. Trocha, M. John, S. Mallal, P. R. Harrigan, et al., J. Virol. **83**, 3407 (2009).

54. X. Jin, D. E. Bauer, S. E. Tuttleton, S. Lewin, A. Gettie, J. Blanchard, C. E. Irwin, J. T. Safrit, J. Mittler, L. Weinberger, et al., J. Exp. Med. **189**, 991 (1999).

55. J. E. Schmitz, M. J. Kuroda, S. Santra, V. G. Sasseville, M. A. Simon, M. A. Lifton, P. Racz, K. Tenner-Racz, M. Dalesandro, B. J. Scallon, et al., Science **283**, 857 (1999).

56. P. Borrow, H. Lewicki, B. H. Hahn, G. M. Shaw, and M. B. Oldstone, J. Virol. **68**, 6103 (1994).

57. R. A. Koup, J. T. Safrit, Y. Cao, C. A. Andrews, G. McLeod, W. Borkowsky, C. Farthing, and D. D. Ho, J. Virol. **68**, 4650 (1994).

58. T. M. Allen, M. Altfeld, S. C. Geer, E. T. Kalife, C. Moore, K. M. O'Sullivan, I. DeSouza, M. E. Feeney, R. L. Eldridge, E. L. Maier, et al., J. Virol. **79**, 13239 (2005).

59. T. M. Allen, X. G. Yu, E. T. Kalife, L. L. Reyor, M. Lichterfeld, M. John, M. Cheng, R. L. Allgaier, S. Mui, N. Frahm, et al., J. Virol. **79**, 12952 (2005).

60. V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania,

T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine, et al., Proc. Natl. Acad. Sci. USA **108**, 11530 (2011).

61. J. Bhalerao and A. M. Bowcock, Hum. Mol. Genet. **7**, 1537 (1998).

62. D. Chessman, L. Kostenko, T. Lethborg, A. W. Purcell, N. A. Williamson, Z. Chen, L. Kjer-Nielsen, N. A. Mifsud, B. D. Tait, R. Holdsworth, et al., Immunity **28**, 822 (2008).

63. P. Bowness, Rheumatology **41**, 857 (2002).

64. B. Peters, J. Sidney, P. Bourne, H. H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, et al., PLoS Biol. **3**, e91 (2005).

65. T. J. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, et al., Nucleic Acids Res. **37**, D690 (2009).

66. E. L. Turnbull, A. R. Lopes, N. A. Jones, D. Cornforth, P. Newton, D. Aldam, P. Pellegrino, J. Turner, I. Williams, C. M. Wilson, et al., J. Immunol. **176**, 6130 (2006).

67. G. M. Gillespie, R. Kaul, T. Dong, H. B. Yang, T. Rostron, J. J. Bwayo, P. Kiama, T. Peto, F. A. Plummer, A. J. McMichael, et al., AIDS **16**, 961 (2002).

68. X. G. Yu, M. Lichterfeld, S. Chetty, K. L. Williams, S. K. Mui, T. Miura, N. Frahm, M. E. Feeney, Y. Tang, F. Pereyra, et al., J. Virol. **81**, 1619 (2007).

69. H. Chen, Z. M. Ndhlovu, D. Liu, L. C. Porter, J. W. Fang, S. Darko, M. A. Brockman, T. Miura, Z. L. Brumme, A. Schneidewind, et al., Nat. Immunol. **13**, 691 (2012).

70. C. L. Thio, X. Gao, J. J. Goedert, D. Vlahov, K. E. Nelson, M. W. Hilgartner,

S. J. O'Brien, P. Karacki, J. Astemborski, M. Carrington, et al., J. Virol. **76**, 4792 (2002).

71. S. M. McKiernan, R. Hagan, M. Curry, G. S. McDonald, A. Kelly, N. Nolan, A. Walsh, J. Hegarty, E. Lawlor, and D. Kelleher, Hepatology **40**, 108 (2004).

72. A. Košmrlj, J. Stat. Phys. **142**, 1277 (2011).

Figure 1: T cell recognition of signatures of pathogens. (a) Antigen-presenting cells (APCs) engulf pathogens and process their proteins into short peptides, which are bound to major histocompatibility complex (MHC) proteins and presented on the surface. T cell receptors (TCRs) bind to peptide-MHCs, and sufficiently strong binding enables intracellular signaling and gene transcription, leading to T cell activation. APCs also present self-peptides derived from self-proteins, but typically T cells are not activated by them. (b) TCR recognition of pathogen-derived pMHC molecules is both highly specific and degenerate. It is specific because if a TCR recognizes (black check-mark) a peptide (green), most point mutations of the peptide's amino acids (red) abrogate recognition (red cross). However, a given TCR can also recognize diverse peptide sequences (green, blue, yellow). Panel (a) is adapted from Figure 1a in Ref. [11].

Figure 2: Immature T cells (thymocytes) develop in the thymus. Thymocytes migrate through the thymus and interact with diverse self peptide–major histocompatibility complexes (self-pMHCs) presented on the surface of thymic antigen presenting cells (APCs). A T cell's receptor (TCR) must bind to at least one of these self-pMHCs moderately to exit the thymus and become a part of the individual's T cell repertoire (positive selection). A T cell with a TCR that binds to any self-pMHC with an affinity that exceeds a sharply defined threshold dies in the thymus (negative selection). This figure is adapted from Figure 3a in Ref. [11].

Figure 3: Effects of thymic selection on the characteristics of TCRs selected against M types of self peptides. (a) Schematic representation of the interface between TCR and pMHCs. The region of the TCR contacting the peptide is highly variable and is modeled by a string of amino acids of length $N \sim 5$. The peptide is also treated similarly. The binding free energy between the TCR and the entire pMHC is computed as described in the text. (b) Amino acid composition of selected TCRs. TCRs selected against many types of self-peptides in the thymus have peptide contact residues that are enriched in amino acids that interact weakly with other amino acids. (c) Probability density distribution of E_c values (strength of TCR binding to MHC) of TCRs selected against M types of self peptides. TCRs selected against many types of self peptides are more likely to bind weakly to MHC. (d) Amino acid composition of pathogenic peptides that are recognized by at least one of the selected TCRs. TCRs selected against many types of self peptides recognize only pathogenic peptides that are enriched with strongly interacting amino acids. Amino acids on the abscissa in (b) and (d) are ordered according to their largest interaction strength with other amino acids in the interaction matrix, J. This figure is adapted from Figs. 1 and 4 in Ref. [17].

Figure 4: Mechanism for specificity and degeneracy of TCR recognition of antigenic peptides: (a) Histogram of the number of important contacts (defined in text) with which T cells recognize pathogenic peptides. T cells selected against many self peptides recognize pathogenic peptides via many important contacts and are thus specific. In contrast, T cells selected against few types of self peptides recognize pathogenic peptides with only a few important contacts and are thus cross-reactive. (b) The weakly interacting amino acids (brown) on the TCR bind to strongly interacting amino acids (red, blue) on antigenic peptides resulting in multiple moderate scale interactions that add up to a total binding free energy that is large enough for recognition. Because antigen recognition is mediated by multiple interactions of moderate value, each contact makes a significant contribution to the total interaction free energy necessary for recognition. Therefore, disrupting any interaction by mutating one of the strongly interacting amino acids on the peptide results (shown as a change from red to yellow color) in abrogation of recognition. At the same time TCR recognition of antigenic peptides is degenerate, because there are many combinatorial ways of distributing strongly interacting amino acids (red, blue) along the peptide, which results in a sufficiently strong binding with TCR for recognition. This figure is adapted from Fig. 3d in Ref. [11] and Fig. 4a in Ref. [17].

Figure 5: Representation of the dependency of the parameters (a) β, a measure of amino acid composition of selected TCRs, and (b) β^*, a measure of amino acid composition of pathogenic-peptides recognized by selected TCRs, on the number of selected TCRs (K), the number of self peptides (M) against which TCRs were selected, and the threshold for negative selection E_n. In (a) the region between the black lines corresponds to $\beta = 0$, to the right (left) of which negative (positive) selection is dominant, and weak (strong) amino acids are selected. The blue dashed lines in (a) and (b) indicate the relevant parameter values for thymic selection in mouse. In (b) solid black lines separate regions with $\beta^* > 0$ (only foreign peptides with strongly interacting amino acids are recognized) and $\beta^* = 0$ (every foreign peptide is recognized). The region below the black dashed line in (b) correspond to $\beta = 0$ (every TCR is selected). In (b) the threshold for negative selection (E_n) is fixed. This figure is adapted from Figs. 2 and 4 in Ref. [17].

Figure 6: HLA-B alleles associated with greater ability to control HIV correlate with smaller self-peptide binding propensities. A large group of HIV infected people were divided into a controller cohort (low levels of HIV RNA) and a progressor cohort (high levels of HIV RNA). The odds ratio is defined as $\frac{p_w/p_{wo}}{c_w/c_{wo}}$, where p_w and p_{wo} (c_w and c_{wo}) are the numbers of individuals in the progressor cohort (the controller cohort) with and without this HLA, respectively. People with HLA alleles associated with an odds ratio value greater or less than one are more likely to be progressors or controllers, respectively. The fraction of peptides derived from the human proteome that bind to a given HLA allele was determined with predictive bioinformatics algorithms [33]. The error bars represent the 95% confidence intervals for odds ratio. The dotted line corresponds to equal odds for an allele being associated with progressors and controllers. Included are only those HLA-B alleles that were statistically significantly associated with HIV control or progression. This figure is adapted from Fig. 3 in Ref. [33].

Figure 1

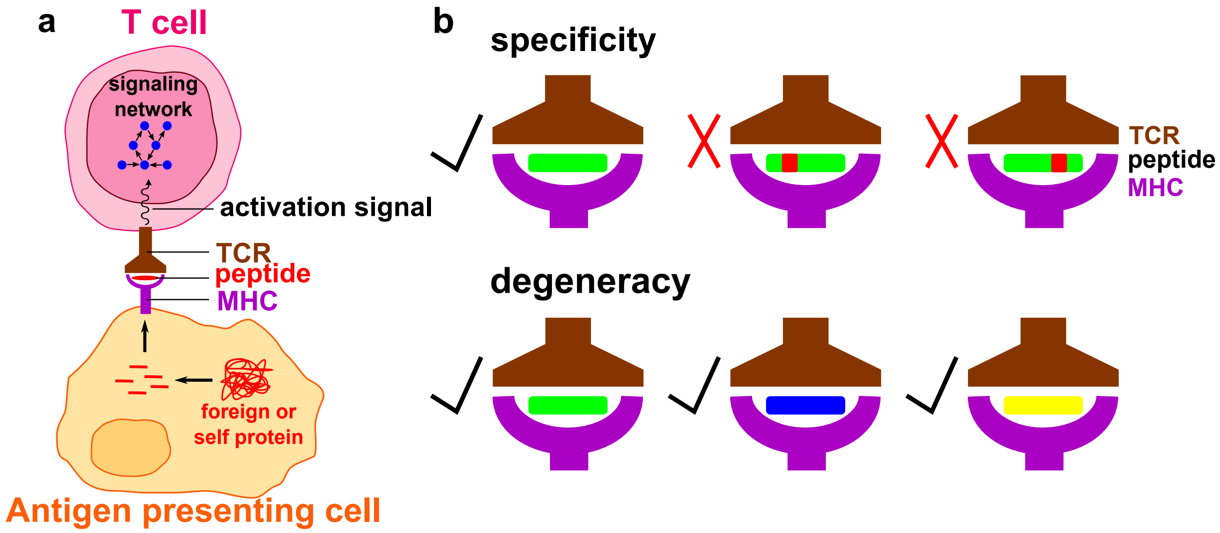

a
T cell
signaling network
activation signal
TCR
peptide
MHC
foreign or self protein
Antigen presenting cell
b
specificity
✓
✗
✗
TCR
peptide
MHC
degeneracy
✓
✓
✓

Figure 2

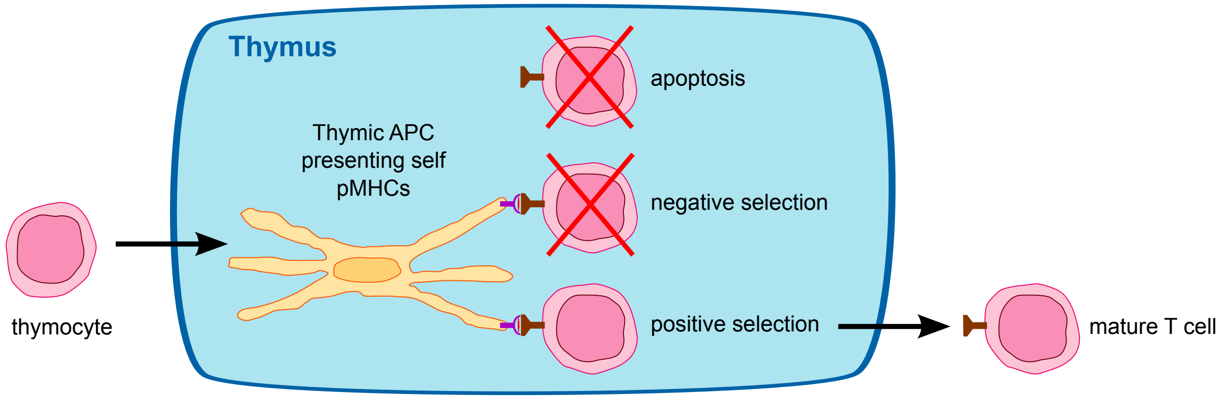
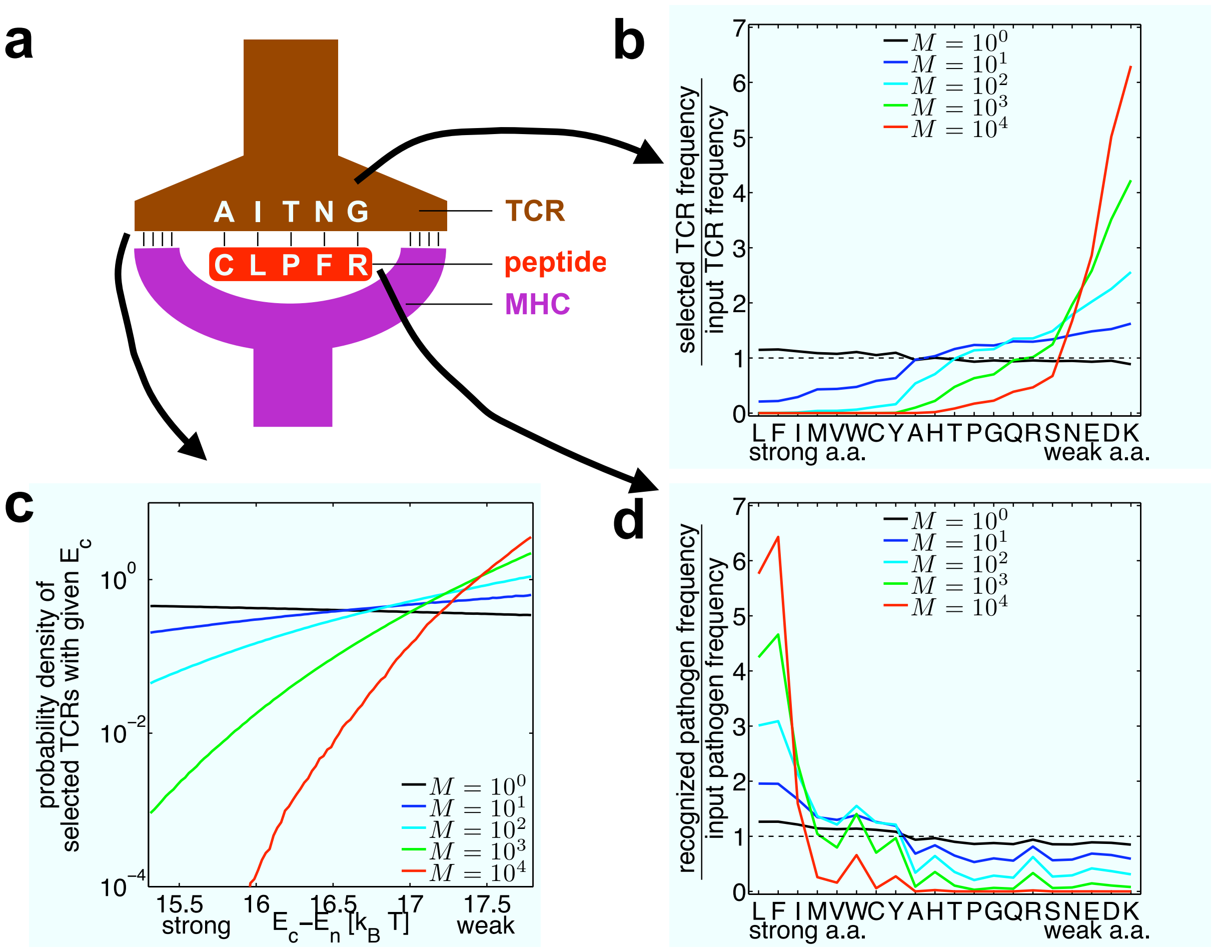

Thymus
apoptosis
Thymic APC
presenting self
pMHCs
negative selection
thymocyte
positive selection
mature T cell

Figure 3



a
A I T N G
C L P F R
TCR
peptide
MHC
b
selected TCR frequency / input TCR frequency
$M = 10^0$
$M = 10^1$
$M = 10^2$
$M = 10^3$
$M = 10^4$
0
1
2
3
4
5
6
7
L F I M V W C Y A H T P G Q R S N E D K
strong a.a.
weak a.a.
c
probability density of selected TCRs with given E_c
10^0
10^{-2}
10^{-4}
$M = 10^0$
$M = 10^1$
$M = 10^2$
$M = 10^3$
$M = 10^4$
15.5
16
16.5
17
17.5
strong
$E_c - E_n$ [k_B T]
weak
d
recognized pathogen frequency / input pathogen frequency
$M = 10^0$
$M = 10^1$
$M = 10^2$
$M = 10^3$
$M = 10^4$
0
1
2
3
4
5
6
7
L F I M V W C Y A H T P G Q R S N E D K
strong a.a.
weak a.a.

Figure 4

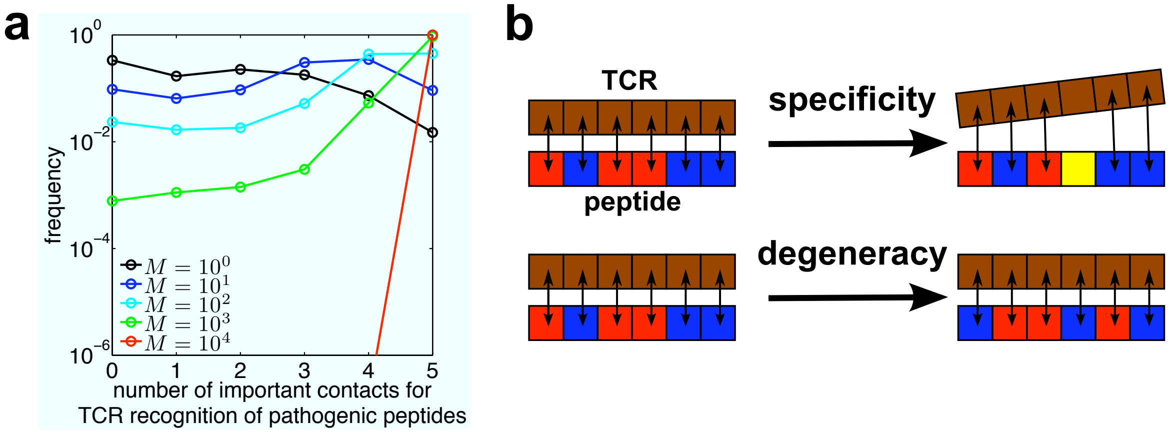

a
10^{0}
10^{-2}
10^{-4}
10^{-6}
frequency
0
1
2
3
4
5
number of important contacts for
TCR recognition of pathogenic peptides
$M = 10^0$
$M = 10^1$
$M = 10^2$
$M = 10^3$
$M = 10^4$
b
TCR
peptide
specificity
degeneracy

Figure 5

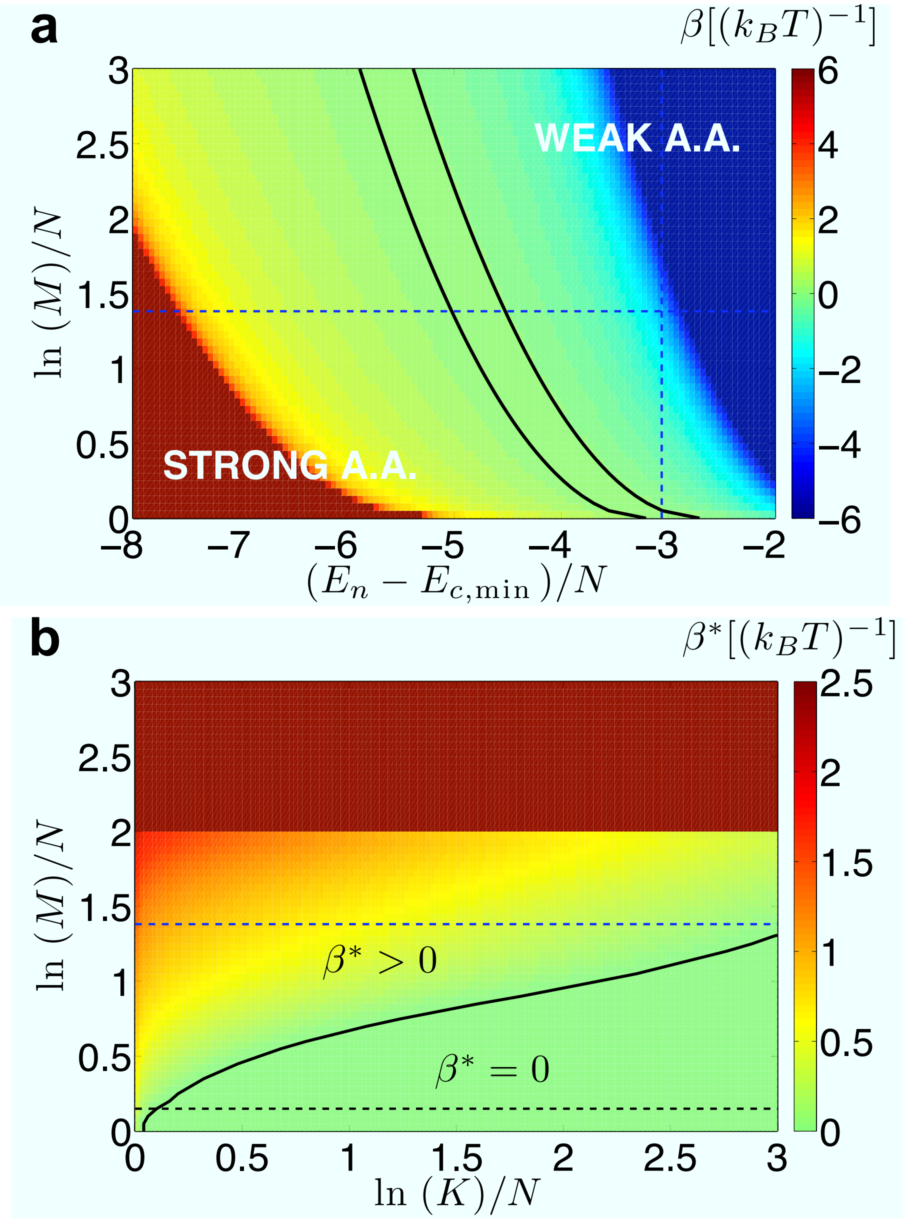
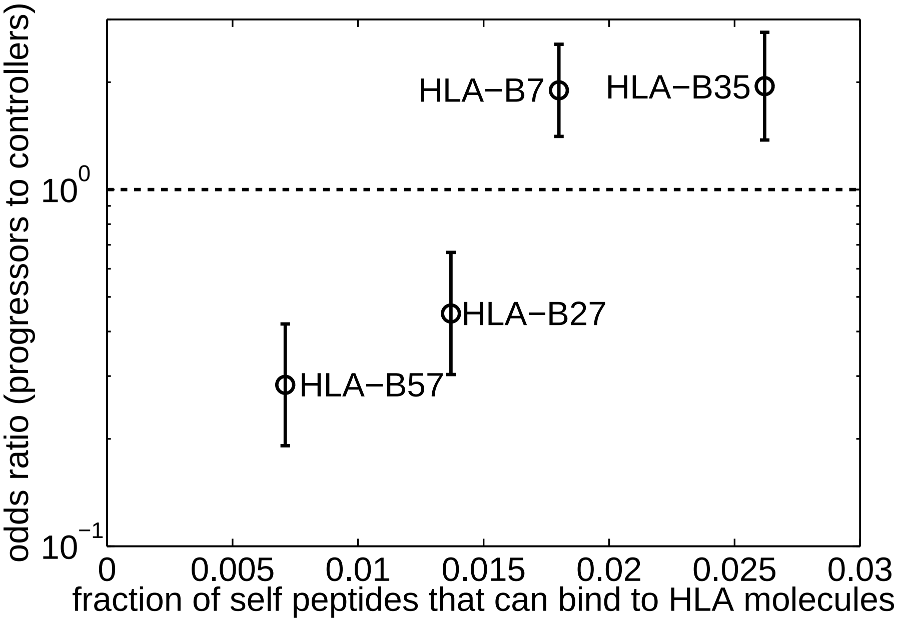

a
$\beta[(k_BT)^{-1}]$
WEAK A.A.
STRONG A.A.
$\ln(M)/N$
$(E_n - E_{c,\min})/N$
3
2.5
2
1.5
1
0.5
0
−8
−7
−6
−5
−4
−3
−2
6
4
2
0
−2
−4
−6
b
$\beta^*[(k_BT)^{-1}]$
$\beta^* > 0$
$\beta^* = 0$
$\ln(M)/N$
$\ln(K)/N$
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
2.5
2
1.5
1
0.5
0

Figure 6


odds ratio (progressors to controllers)
10^0
10^{-1}
HLA−B7
HLA−B35
HLA−B27
HLA−B57
0
0.005
0.01
0.015
0.02
0.025
0.03
fraction of self peptides that can bind to HLA molecules