Journal of Theoretical Medicine, March 2003 Vol. 5 (1), pp. 1–22

Taylor & Francis
healthsciences

Thymic Presentation of Autoantigens and the Efficiency of Negative Selection

HUGO A. VAN DEN BERG*,† and CARMEN MOLINA-PARÍS‡,§

Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK

(Received 11 March 2003; In final form 21 June 2003)

Antigen recognition by the adaptive cellular immune system is based on a diverse repertoire of antigen receptors. Since this repertoire is formed by genetic recombination, a number of receptors are autoreactive by chance, giving rise to the threat of autoimmune disease. Potentially autoreactive T lymphocytes (T cells) are rendered ineffective by various tolerance mechanisms. One of these mechanisms is negative selection, the deletion from the repertoire of immature autoreactive T cells in the thymus. The present paper shows how to assess the contribution made by negative selection relative to other tolerisation mechanisms by deducing the impact of negative selection on the T cell repertoire from the statistics of autoantigen presentation in the thymus.

Keywords: T cell tolerance; Negative selection; Autoimmunity; Large deviations theory

INTRODUCTION

T cell tolerance is a complex phenomenon which comprises both central and peripheral tolerance: lymphocytes acquire the former while they are developing in the primary lymphoid tissues, the latter as mature cells recirculating through the secondary lymphoid tissues (Janeway and Travers, 1997). The present paper shows how to resolve T cell tolerance into its central and peripheral contributions by a comparative analysis of the statistics of autoantigen presentation in the primary and secondary lymphoid tissues.

The adaptive cellular immune system recognises pathogenic antigens by means of the T cell antigen receptor (TCR), which interacts with peptide antigens displayed on the surface of antigen presenting cells (APCs) by glycoproteins belonging to the major histocompatibility complex (MHC) (Janeway and Travers, 1997). The immune system contains millions of distinct TCR molecules, formed by random rearrangement of the gene segments encoding the region of the TCR molecule that interacts with the peptide-MHC (pMHC) complex (Janeway and Travers, 1997; Arstila *et al.*, 1999). Each T cell expresses one specific TCR species or clonotype, unique to the T cell and the clone to which it belongs.

Since the generation of TCR clonotypes proceeds at random, some of them inevitably are autoreactive, that is, their TCR molecule recognises one or more antigens derived from the body's own proteins (autoantigens). Such autoreactive clones are kept in check by various tolerisation mechanisms which act to prevent their activation and concomitant autoimmune disease. Following recognition of autoantigens presented by a tolerising APC, an autoreactive T cell may be eliminated, or forced to reduce its responsiveness, or to become a suppressor cell (Webb *et al.*, 1990; Antonia *et al.*, 1995; Smith *et al.*, 1997; Seddon and Mason, 1999; Roncarolo and Levings, 2000; Steinman *et al.*, 2000; Hawiger *et al.*, 2001).

The T cell repertoire undergoes central tolerisation in the thymus, where numerous immature T cells (thymocytes) that recognise autoantigens are induced to undergo apoptosis (Kappler *et al.*, 1987; Kisielow *et al.*, 1988; Surh and Sprent, 1994). While this process of negative selection is generally thought to prevent maturation of many autoreactive T cells, it seems likely that some degree of autoreactivity remains in the mature repertoire (Tanchot *et al.*, 1997; Bouneaud *et al.*, 2000; Visser *et al.*, 2000; Zippelius *et al.*, 2002). Low-avidity TCR interaction with autoantigens accounts for part of this residual autoreactivity (Bouneaud *et al.*, 2000; Hernández *et al.*, 2000; Nugent *et al.*, 2000; Visser *et al.*, 2000).

*Corresponding author. Tel.: +44-2476-523698. Fax: +44-2476-524182. E-mail: hugo@maths.warwick.ac.uk
†Supported by the Wellcome Trust.
‡Supported by the EPSRC.
§Present address: Department of Applied Mathematics, University of Leeds, Leeds LS2 9JT.

ISSN 1027-3662 print/ISSN 1607-8578 online © 2003 Taylor & Francis Ltd
DOI: 10.1080/10273662031000160491O

Moreover, negative selection would fail to abolish autoreactivity completely if negatively selecting cells in the thymus could only present constitutive and lineage-specific peptides, as opposed to tissue-specific and sequestered peptides, a traditional argument no longer deemed to be strictly correct (Klein *et al.*, 1998; Sospreda *et al.*, 1998).

A deeper reason why autoreactivity cannot be completely eliminated is that it becomes virtually impossible for negative selection to abolish autoreactivity entirely if the number of autoantigens recognised by a pre-selection thymocyte is much larger than 1, which is likely to be the case (Hogquist *et al.*, 1997; Mason, 2001). Some degree of residual autoreactivity in the mature T cell repertoire thus seems inevitable, which suggests that negative selection should be regarded as modifying rather than eliminating autoreactivity in the TCR repertoire.

The autoreactivity of an individual TCR clonotype involves more than merely the number of autoantigens recognised by its TCR molecule. Equally important for the immunogenic potential of an autoantigen are (i) its ubiquity, that is, the frequency with which a mature recirculating T cell will encounter the autoantigen as it visits secondary lymphoid tissues throughout the body (Janeway and Travers, 1997; Klein *et al.*, 1998) and (ii) its presentation level, that is, its copy number on the surface of the APCs (Akkaraju *et al.*, 1997; Byers and Lindahl, 1999; Kurts *et al.*, 1999; Morgan *et al.*, 1999; Sebza *et al.*, 1999; Reay *et al.*, 2000). Moreover, a thymocyte need not encounter a given autoantigen with the same ubiquity or presentation level as a mature T cell does. For instance, thymic negatively selecting cells may not be able to present certain autoantigens derived from proteins with non-constitutive (inducible) or tissue-restricted expression, and these autoantigens would have zero ubiquity in the thymus. However, a growing body of experimental evidence indicates to the contrary that negatively selecting cells are able to express and present rare peripheral antigens that are restricted to specific cell types in the periphery (Antonia *et al.*, 1995; Fritz and Zhao, 1996; Smith *et al.*, 1997; Farr and Rudensky, 1998; Hanahan, 1998; Klein *et al.*, 1998; Sospreda *et al.*, 1998). Such ectopic thymic expression of peripheral autoantigens would seem to broaden considerably the scope and efficiency of central tolerance.

The analysis of the autoreactivity of the T cell repertoire is technically challenging because of the presence of two kinds of variability: interclonotypic variability due to random differences between the autoreactivity of various TCR clonotypes and intraclonotypic variability due to fluctuations in the autoantigens presented on various APCs. In a sense, the repertoire is a "distribution of distributions": each clonotype has its own distribution law from which it samples as it registers signals on subsequent APCs, and then the whole repertoire is a collection of samples from a distribution of laws. Tolerance can only affect the latter distribution, not the law itself, since it is intrinsic to the TCR's molecular structure.

An earlier paper (van den Berg *et al.*, 2001) analysed the simple extreme case where self peptides encountered on APCs in the secondary lymphoid tissues are either constitutive (present on all APCs) or exceedingly rare (found on a very small fraction of APCs). All clonotypes then share the same law up to an additive term, and the TCR repertoire is fully characterised by the distribution of the latter term over the clones. However, as noted in that paper, on this simplification one cannot deal with autoantigens of intermediate ubiquity, differences in self presentation statistics in the thymus as opposed to the periphery, and the distinct contributions made, respectively, by central and peripheral tolerance. The purpose of the present paper is to carry out a more general analysis which allows these issues to be addressed. Thus, the theory developed in this paper describes precisely how patterns of thymic presentation shape the statistical structure of the mature T cell repertoire. In particular, the present theory shows how data on the thymic ubiquities and presentation levels of autoantigens, ranging from very rare to constitutive, can be used to characterise how negative selection modifies the repertoire, which will enable immunologists to delineate the relative contributions of central versus peripheral tolerance.

Organisation of the paper: The second section develops a large deviations description of the stochastic variability of T cell stimulation due to autoantigens, broken down into interclonotypic variability due to random differences between the autoreactivity of various TCR clonotypes and intraclonotypic variability due to fluctuations in the autoantigens presented on various APCs. The third section presents the main results. The repertoire structure induced by central tolerance is related explicitly to thymic autoantigen presentation statistics. Furthermore, the efficiency of negative selection is precisely characterised. The fourth section outlines an application of the theory, showing by means of two examples how comparison of central and peripheral autoantigen presentation can be used to elucidate the role of negative selection. Notation used throughout is summarised in Table I; additional notation is always explained locally.

FLUCTUATIONS IN T CELL STIMULATION BY SELF

Upon conjugation with an antigen-presenting cell, a T cell registers a signal through its TCR molecules due to the peptides presented on the MHC molecules on the APC surface. This TCR signal will be represented as a weighed sum over the contributions due to the various pMHC species, with weighing factors corresponding to the densities of the various pMHC species on the APC. The T cell will be assumed to respond when the TCR signal exceeds a cellular threshold (Viola and Lanzavecchia, 1996).

TABLE I Notation

Symbol	Interpretation
i	index for T cell clonotypes
j	index for pMHC species
k	index for self presentation components
N	number of self pMHC species with which a given TCR can interact, with or without productive recognition
N_k	number of self pMHC species belonging to component k
$\hat{N}_{ik}$	number of self pMHC species recognized by clonotype i, belonging to component k
$\hat{\boldsymbol{L}}_i$	clonotype law, $\hat{\boldsymbol{L}}_i = \{\hat{L}_{ik}\}_{k=1}^K$ where $\hat{L}_{ik} = \hat{N}_{ik}/N$ is the autorecognition frequency of clone i for component k
$\mathscr{L}_N$	set of all possible clonotype laws
K	number of self presentation components
$\mathscr{K}_k$	$j \in \mathscr{K}_k$ iff pMHC species j belongs to component k; $\lvert\mathscr{K}_k\rvert = N_k$
μ	probability that a random TCR chosen recognizes a pMHC species that is chosen at random from among the pMHCs with which the TCR can interact
$\boldsymbol{\zeta}$	APP, $\boldsymbol{\zeta} = \{\zeta_j\}_{j=1}^N$ where ζ_j is relative presentation level of pMHC species j
$\bar{\zeta}_j$	relative presentation level of pMHC species j, averaged over APPs
$I_{j\boldsymbol{\zeta}}$	indicator which equals 1 when $\zeta_j > 0$ in APP $\boldsymbol{\zeta}$, and 0 otherwise
u_j	ubiquity, probability that pMHC species j appears in a randomly chosen APP $\boldsymbol{\zeta}$ $u_j = \mathbb{P}(I_{j\boldsymbol{\zeta}} = 1)$
n_{div}	presentation diversity, $n_{\mathrm{div}} = 1/\sum_{j=1}^N \bar{\zeta}_j^2$
$n_{k\boldsymbol{\zeta}}$	number of self pMHC species with $\zeta_j > 0$ in APP $\boldsymbol{\zeta}$, belonging to component k
$\hat{n}_{ik\boldsymbol{\zeta}}$	number of self pMHC species with $\zeta_j > 0$ in APP $\boldsymbol{\zeta}$, belonging to component k and recognized by clonotype i
$w_{i\boldsymbol{\zeta}}$	TCR signal elicited in clonotype i by APP $\boldsymbol{\zeta}$
w_{ij}	TCR signal elicited in clonotype i by a pMHC complex of species j
w_{act}	cellular activation threshold
w_{thy}	thymic negative selection threshold
$\bar{w}_i$	TCR signal due to self in clonotype i, averaged over APPs
σ^2_{APP}	variance over APPs of TCR signal due to self in clonotype i
$\boldsymbol{\pi}$	component partitioning, $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ where $\pi_k = N_k/N$
$\boldsymbol{\rho}$	component presentation, $\boldsymbol{\rho} = \{\rho_k\}_{k=1}^K$ where $\rho_k = \lim_{N\to\infty} \sum_{j\in\mathscr{K}_k} \zeta_j$
$\boldsymbol{u}$	component ubiquities, $\boldsymbol{u} = \{u_k\}_{k=1}^K$
$\bar{u}$	mean effective ubiquity, $\bar{u} = \left[\sum_{k=1}^K \rho_k^2 \pi_k^{-1}\right] \Big/ \left[\sum_{k=1}^K \rho_k^2 (\pi_k u_k)^{-1}\right]$
r_j	presentation propensity of pMHC species j
$r_{T\boldsymbol{\zeta}}$	total presentation propensity, $r_{T\boldsymbol{\zeta}} = \sum_j^N r_j \boldsymbol{I}_{j\boldsymbol{\zeta}}$
m	number of rounds of negative selection in the thymus
m_{crit}	critical number of rounds of negative selection, $m_{\mathrm{crit}} = (1-\bar{u})/\bar{u}$
P_{surv}	probability that a thymocyte survives negative selection
$\boldsymbol{I}, I, J$	large deviations rate functions

APP, antigen presentation profile; TCR, T cell antigen receptor; pMHC, major histocompatibility complex molecule, presenting a peptide; pAPC, professional antigen-presenting cell.
A prime indicates statistics for peripheral pAPCs; unprimed statistics refer to cells effecting negative selection. Not included are notations used and explained locally. See "Discussion" section for further remarks on the interpretation of P_{surv}.

Autorecognition of various autoantigens by a given clonotype can no longer be treated as statistically independent following negative selection: if A and B are two arbitrary autoantigens, the probability that a mature T cell, chosen at random, recognises A given that it recognises B is lower than the unconditional probability that a random mature T cell recognises A (both probabilities are of course generally lower than the probability that A is recognised by a thymocyte chosen at random prior to negative selection). Since the total number of peripheral autoantigens, rare as well as abundant, is very large, a direct attack on the correlations induced by negative selection using iterative conditioning quickly becomes very cumbersome. Fortunately, a more tractable way to deal with the correlations is offered by large deviations techniques, which exploit the vastness of the number of autoantigens.

Variability of the TCR signal due to recognition of autoantigens has two major sources: one is interclonotypic variability, which arises because different TCR molecules recognise different antigens, and another is intraclonotypic variability due to random fluctuations in antigen presentation. To represent both kinds of fluctuations in a manner that is amenable to large deviations techniques, the autoantigens will be lumped into *presentation components.*

T Cell Activation

The TCR signal not only depends on the TCR clonotype i, but also on the copy numbers of the various pMHC species on the APC. The latter can be represented as an antigen presentation profile (APP)

$$\zeta \stackrel{\mathrm{def}}{=} \{\zeta_1, \zeta_2, \ldots\}$$

where ζ_j denotes the relative presentation level of pMHC species j, defined by

$$\zeta_j \stackrel{\mathrm{def}}{=} Z_j/M_T$$

where M_T is the number of pMHC molecules capable of binding to the TCR at hand and Z_j is the number of pMHC molecules of species j; the relative presentation level ζ_j sums to 1 over the pMHC species j.

Let $w_{i\zeta}$ denote the TCR signal registered by a T cell of TCR clonotype i conjugated with an APC presenting

a pMHC ensemble $\boldsymbol{\zeta}$. The *summation hypothesis* states that $w_{i\boldsymbol{\zeta}}$ can be represented as follows:

$$w_{i\boldsymbol{\zeta}} = \sum_{j=1}^{N} \zeta_j w_{ij} \tag{1}$$

where N denotes the total number of self pMHC species that can bind to the TCR at hand (this number is assumed to be the same for all TCR clonotypes). In the present paper it will be assumed that w_{ij} is a Bernoulli variate:

$$w_{ij} = \begin{cases} 0 \text{ when TCR } i \text{ does not productively recognise pMHC species } j \\ 1 \text{ when TCR } i \text{ productively recognises pMHC species } j. \end{cases} \tag{2}$$

For a given clonotype i the values of w_{ij} for the various pMHC species j determine the identity of the clonotype.

The restriction of w_{ij} to two values (0 and 1) is motivated in Appendix A. The quantities $w_{i\boldsymbol{\zeta}}$ and w_{ij} are dimensionless, and have been scaled relative to the maximum value attainable by the TCR signal (see Appendix A).

Let μ denote the probability that a randomly selected TCR recognises a randomly selected pMHC species; thus, $\mu = \mathbb{P}\{w_{ij} = 1\}$ for a randomly chosen pair (i,j). The parameter μ represents the inherent degeneracy of TCR/pMHC recognition (Mason, 1998; Borghans *et al.*, 1999). Its reciprocal $1/\mu$ is the inherent specificity, and equals the average number of clonotypes that must be examined before one is found that recognises the pMHC species at hand. Pathogenic and normal autoantigens are not formally distinguished: the same parameter μ denotes both the probability that a T cell will recognise a foreign antigen and the probability that an autoantigen will be recognised by a positively selected thymocyte before negative selection. It may be assumed that $\mu \ll 1$: even though TCR recognition degeneracy is considerable (Mason, 1998), any given TCR will be unable to interact productively with the vast majority of pMHC species with which it can interact; μ is typically estimated to lie in the range 10^{-5}–10^{-4} (Gavin and Bevan, 1995; Butz and Bevan, 1998; Mason, 1998); for certain antigens the frequency of T cells in the naïve repertoire that recognise the antigen may be higher: for instance, it was found that over 1 in 2500 $CD8^+$ T cells are specific for the Melan-A/MART-1 autoantigen in HLA-A2 individuals (Pittet *et al.*, 1999), and the frequency of specific precursors in the repertoire to antigens such as cytomegalovirus may be higher still (Oelke *et al.*, 2003).

The *threshold hypothesis* states that the T cell becomes activated when the TCR signal is greater than some threshold value. In fact, the T cell may be capable of various responses, some of which may be more readily evoked than others (Valitutti *et al.*, 1996; Itoh and Germain, 1997). Accordingly, the T cell may be assumed to have various different threshold values, each corresponding to a particular response. Two important examples of such responses are (i) the naïve T cell's decision to commit to differentiation and proliferation, which happens when $w_{i\boldsymbol{\zeta}}$ exceeds w_{act} (Viola and Lanzavecchia, 1996; Lanzavecchia and Sallusto, 2000) and (ii) the thymocyte's entry into apoptosis, which happens when $w_{i\boldsymbol{\zeta}}$ exceeds the selection threshold w_{thy} (Bouneaud *et al.*, 2000; Savage and Davis, 2001).

Component Representation of Self

The number N of self pMHC species that can interact with a given TCR is very large (Bevan *et al.*, 1994; Mason, 2001), and this fact can be exploited to give large deviations estimates for the repertoire structure. Thus, it is essential that antigen presentation be described in a way that allows the limit $N \to \infty$ to be taken. To this end, the self pMHC species will be partitioned into $K < \infty$ self-presentation components, such that all pMHC species belonging to a given component have two characteristics in common: their frequency of occurring in an APP and their typical presentation level.

The component approach allows a compact representation of fluctuations in TCR signalling due to self. The essential idea is that each component collects self pMHC species that share their ubiquity and mean presentation level. This section first explains the general concept, and then details the particular model for self-fluctuations used in the present paper.

The Concept of Self Presentation Components

A given T cell will in general not register the same TCR signal during subsequent encounters with different APCs, because the ensemble of presented pMHC species will differ; if the encounters are sufficiently far apart in time, the same will even be true of the same APC. Such variations in the APP occur because not every self peptide occurs in every APP, even within a defined class of APCs such as the negatively-selecting cells. Moreover, the presentation level of a peptide will vary as the total numbers of peptides present at non-zero presentation levels also fluctuates between APPs. The idea behind the component approach introduced in this paper is to characterise these various kinds of fluctuations in a non-degenerate manner as $N \to \infty$, by representing the TCR signal $w_{i\boldsymbol{\zeta}}$ in terms of presentation components rather than the pMHC species themselves.

Every self peptide species j has a specific ubiquity u_j relative to a given class of APCs: $u_j \in [0,1]$ represents the probability that the peptide is presented by a randomly chosen APC from this class. A pMHC species present in every APP will have ubiquity one, while

a species presented very rarely will have a ubiquity close to zero. Finally, a peptide never expressed by the APCs at hand will have ubiquity zero for that class of APCs. The number N refers to all autoantigens that can interact with a given TCR, with or without productive recognition. One would therefore expect many ubiquities to be identically zero for any particular class of APCs, although professional APCs may be more versatile than many other cell types in this respect (Steinman *et al.*, 2000). Besides their ubiquities, pMHCs also differ with respect to their presentation levels (Hunt *et al.*, 1992; Morgan *et al.*, 1999; Reay *et al.*, 2000). Thus, every self peptide is furthermore characterised by its average presentation level $\bar{\zeta}_j$. A pMHC species j is thus characterised by two parameters: its ubiquity u_j and its average relative presentation level $\bar{\zeta}_j$.

The component approach distributes the N self peptides among K components so that peptides that belong to the same component have the same ubiquity and average presentation level. Thus, these parameters can be indexed by components k instead of peptides j, such that $u_j = u_k$ and $\bar{\zeta}_j = \bar{\zeta}_k$ whenever pMHC species j belongs to component k. While a pMHC species will generally have to be assigned to a component whose $(u_k, \bar{\zeta}_k)$ does not exactly match $(u_j, \bar{\zeta}_j)$, the number of components K can always be increased for a finer-grained classification.

Two classes of professional antigen-presenting cells will be considered in this paper: the negatively selecting cells which mediate negative selection and the professional antigen-presenting cells (pAPCs) in the secondary lymphoid tissues which activate naïve T cells. The latter will be distinguished by a prime: u'_k, $\bar{\zeta}'_k$. For two peptides to belong to the same self presentation component, it is only required that they share ubiquity and mean presentation level in both classes, while it is not required that these values be the same in the two classes: where k denotes a component and j_1 and j_2 two pMHC species,

$$\begin{aligned} &j_1 \in k \quad \text{and} \quad j_2 \in k \quad \text{iff} \\ &u_{j_1} = u_{j_2} = u_k \quad \& \quad \bar{\zeta}_{j_1} = \bar{\zeta}_{j_2} = \bar{\zeta}_k \quad \& \\ &u'_{j_1} = u'_{j_2} = u'_k \quad \& \quad \bar{\zeta}'_{j_1} = \bar{\zeta}'_{j_2} = \bar{\zeta}'_{\mathrm{k}} \end{aligned}$$

while not necessarily $u_k = u'_k$ or $\bar{\zeta}'_k = \bar{\zeta}'_k$. Consequently, two distinct components k and ℓ may have the thymic statistics in common ($u_k = u_\ell$ and $\bar{\zeta}_k = \bar{\zeta}_\ell$) but owe their status as distinct components to peripheral differences (e.g. $u'_k \neq u'_\ell$). The ubiquities for the various components are collected in a K-vector $\boldsymbol{u}$ (for the class of negatively selecting cells) or $\boldsymbol{u}'$ (for the class of pAPCs in the secondary lymphoid tissues).

In what follows it will be useful to focus on the fraction of MHC molecules that present a pMHC belonging to component k. This fraction is the component presentation level ρ_k:

$$\rho_k \stackrel{\text{def}}{=} \lim_{N\to\infty} \sum_{j\in\mathscr{K}_k} \zeta_j \tag{3}$$

where $\mathscr{K}_k$ is a subset of $\{1,2,\ldots,N\}$ such that $j \in \mathscr{K}_k$ iff j belongs to component k. The component presentation levels are collected in a K-vector $\boldsymbol{\rho}$ with $\sum_{k=1}^{K} \rho_k = 1$; inner products of K-vectors will often be written more compactly, thus: $\langle \mathbf{1}, \boldsymbol{\rho} \rangle \equiv \sum_{k=1}^{K} \rho_k$. The class of negatively selecting thymic stroma cells is then parametrised by ubiquities $\boldsymbol{u}$ and component presentation levels $\boldsymbol{\rho}$, which together describe the statistical fluctuations of self antigen presentation within that class. Similarly, the class of pAPCs in the secondary lymphoid tissues is characterised by its own pMHC ubiquities $\boldsymbol{u}'$ and component presentation levels $\boldsymbol{\rho}'$.

The Ubiquity–Propensity Model

The concept of *presentation propensity* was introduced by van den Berg *et al.* (2001) to describe APP variations between APCs with varying protein expression patterns. The relative presentation level of a peptide depends on a number of factors: for MHC class I presentation, these factors include whether the peptide is expressed in the APC; the level at which it is expressed (Jardetzky *et al.*, 1991); the likelihood that the peptide is properly processed (protein degradation in the proteasome, transport to the intracellular compartment containing the MHC molecules (Stevanović and Schild, 1999)) and its affinity for the MHC binding cleft, which determines its ability to compete with other peptides. The propensity model combines these factors into a single, aggregate propensity measure $r \geq 0$. The key idea is that the ratio between presentation levels of every pair of expressed peptides j and j' is given by the ratio of their propensities, $\zeta_j/\zeta_{j'} = r_j/r_{j'}$ where r_j and $r_{j'}$ are the propensities of the two peptides. Thus, peptides with equal propensities have equal presentation levels, when both are expressed. Class II presentation is modelled in a similar way, although the propensity now reflects the properties of the endocytic processing pathway, and the presence of a peptide in the APP is governed by the protein content of the endocytic vesicles rather than expression by the APC itself.

On the propensity model, the relative presentation level ζ_j of pMHC species j is found by dividing the propensity r_j of the peptide by the sum of the propensities of all peptides that are present in the MHC loading compartment. If $I_{j\boldsymbol{\zeta}}$ is an independent Bernoulli variate which takes on the value 1 with probability u_k when j belongs to component k, then

$$\zeta_j = I_{j\boldsymbol{\zeta}} \frac{r_j}{r_{T\boldsymbol{\zeta}}} \quad \text{with} \quad r_{T\boldsymbol{\zeta}} \stackrel{\text{def}}{=} \sum_{j=1}^{N} r_j I_{j\boldsymbol{\zeta}}, \tag{4}$$

and N independent realisations of Bernoulli variates $I_{j\boldsymbol{\zeta}}$—one for each self pMHC species—are required to specify a single APP $\boldsymbol{\zeta}$.

Let π_k denote the probability that a randomly selected self pMHC belongs to component k, so that there are $N_k \stackrel{\text{def}}{=} \pi_k N$ pMHC species in the kth component ($|\mathscr{K}_k| = N_k$), and the distribution of self pMHC species over the components is represented by a K-vector $\boldsymbol{\pi}$ with $\langle \mathbf{1}, \boldsymbol{\pi} \rangle = 1$. From the fact that $\lim_{N\to\infty} r_T\zeta/N = \sum_{k=1}^K r_k u_k \pi_k$ where r_k denotes the propensity shared by all pMHCs $j \in \mathscr{K}_k$, it follows that the component presentation level is given by

$$\rho_k = \frac{r_k u_k \pi_k}{\sum_{\ell=1}^K r_\ell u_\ell \pi_\ell}.$$

The mean presentation level for a pMHC species belonging to component k can now be expressed as

$$\bar{\zeta}_k = \frac{r_k}{N\sum_{\ell=1}^K r_\ell u_\ell \pi_\ell} = \frac{\rho_k}{u_k \pi_k N}, \tag{5}$$

and the TCR signal due to self, Eq. (1), can be rewritten as a sum over the components (with Eqs. (4) and (5)):

$$w_{i\boldsymbol{\zeta}} = \sum_{j=1}^N I_{j\boldsymbol{\zeta}} \frac{r_j}{r_T\zeta} w_{ij} = \sum_{k=1}^K \frac{r_k}{r_T\zeta} \sum_{j\in\mathscr{K}_k} I_{j\boldsymbol{\zeta}} w_{ij} = \sum_{j=1}^K \frac{r_k}{r_T\zeta} \hat{n}_{ik\boldsymbol{\zeta}}$$

where $\hat{n}_{ik\boldsymbol{\zeta}} \stackrel{\text{def}}{=} \sum_{j\in\mathscr{K}_k} I_{j\boldsymbol{\zeta}} w_{ij}$ denotes the number of pMHC species j that belong to component k and are (i) presented in APP $\boldsymbol{\zeta}$ ($I_{j\boldsymbol{\zeta}} = 1$) and (ii) are productively recognised by the TCR of clonotype i ($w_{ij} = 1$). For large N, the TCR signal due to self becomes:

$$\lim_{N\to\infty} w_{i\boldsymbol{\zeta}} = \sum_{k=1}^K \bar{\zeta}_k \hat{n}_{ik\boldsymbol{\zeta}} = \sum_{k=1}^K \frac{\rho_k}{u_k \pi_k} \frac{\hat{n}_{ik\boldsymbol{\zeta}}}{N} \tag{6}$$

which is well-defined as the quantity $\hat{n}_{ik\boldsymbol{\zeta}}/N$ does not become degenerate in the limit $N \to \infty$. Stochastic fluctuations of the recognised pMHC species $\{\hat{n}_{ik\boldsymbol{\zeta}}\}_{k=1}^K$ are partly due to differences between clonotypes i, and partly due to differences between APPs $\boldsymbol{\zeta}$. In other words, there is both an across-clonotype and a within-clonotype contribution to the fluctuations of $\hat{n}_{ik\boldsymbol{\zeta}}$. Negative selection reshapes the T cell population through differential probabilities of survival, but cannot alter the autorecognition properties of any given thymocyte; all that negative selection can do is to delete certain clonotypes.

Fluctuations across and within Clonotypes

The basic technical difficulty in the analysis of T cell tolerance is the fact that the T cell repertoire is a "distribution of distributions": each clonotype has its own across-APP distribution of self stimulation, and the repertoire as a whole is a collection of such distributions.

The component-based approach allows the introduction of a law (distribution) specific for each clonotype. Variability across the APPs for a given clonotype is governed by the law of that particular clonotype. By contrast, variability over the clonotypes is represented by the ensemble of the clonotype laws, which models the statistical structure of the TCR repertoire, and which is modified by negative selection.

Recognition Frequencies: The Clonotype Law

Let $\hat{N}_{ik}$ denote the number of self pMHC species belonging to a given component k that are recognised by a TCR of clonotype i:

$$\hat{N}_{ik} \stackrel{\text{def}}{=} \sum_{j\in\mathscr{K}_k} w_{ij}. \tag{7}$$

While fixed for a given combination of clonotype i and component k, the number $\hat{N}_{ik}$ varies randomly between clonotypes. Prior to negative selection, $\hat{N}_{ik}$ follows a binomial distribution:

$$\mathbb{P}(\hat{N}_{ik} = x) = \binom{N_k}{x} \mu^x (1-\mu)^{N_k - x} \tag{8}$$

provided that recognition of pMHC species is statistically independent. This is a reasonable assumption in view of the fact that degeneracy in TCR/pMHC recognition is polyspecific, with unfocused cross-reactivity (Hagerty and Allen, 1995; Ignatowicz *et al.*, 1997; Mazza *et al.*, 1998; Anderson *et al.*, 2000).

A TCR clonotype i is fully specified by the numbers $\hat{N}_{ik}$ of recognised pMHC species in the various components k. Normalising these numbers to their respective maximum values, one obtains the *clonotype law* $\hat{\boldsymbol{L}}_i$:

$$\hat{\boldsymbol{L}}_i \stackrel{\text{def}}{=} \left\{\hat{N}_{i1}/N_1, \ldots, \hat{N}_{iK}/N_K\right\}$$

which is an element of the set $\mathscr{L}_N$ defined by

$$\mathscr{L}_N \stackrel{\text{def}}{=} \left\{ \boldsymbol{\nu} : \boldsymbol{\nu} = \{\hat{N}_1/N_1, \ldots, \hat{N}_K/N_K\} \right.$$
$$\left. \text{for some } \{\hat{N}_1, \ldots, \hat{N}_K\} \in \mathscr{N} \right\}$$

where $\mathscr{N} \stackrel{\text{def}}{=} \{0, N_1\} \times \cdots \times \{0, N_K\}$. The kth element of the law $\hat{L}_{ik} = \hat{N}_{ik}/N_k$ is the clonotype's autorecognition frequency for pMHC species belonging to component k. Thus, the component approach represents a T cell's autorecognition entirely by the law of the clone to which it belongs. It will often be convenient to refer more briefly to a T cell's law $\hat{\boldsymbol{L}}$.

Since each of the $\hat{N}_{ik}$ recognised autoantigens has a probability u_k of being present at a non-zero presentation level in a random APP, the average TCR signal due to self of clonotype i, the clonotype mean $\bar{w}_i$, is $\sum_{k=1}^K \bar{\zeta}_k u_k \hat{N}_{ik}$, which can be written succinctly in terms of the component presentation levels and the clonotype law:

$$\bar{w}_i = \langle \boldsymbol{\rho}, \hat{\boldsymbol{L}}_i \rangle. \tag{9}$$

Thus, a T cell with autorecognition frequencies $\hat{\boldsymbol{L}}$ experiences an average TCR signal $\langle \boldsymbol{\rho}, \hat{\boldsymbol{L}} \rangle$ from self peptides it encounters during negative selection in the thymus, and $\langle \boldsymbol{\rho}', \hat{\boldsymbol{L}} \rangle$ is the average TCR signal due to autorecognition registered in the secondary lymphoid tissues. A large deviations rate function for $\bar{w}_i$ is derived in Appendix B.1.

Prior to negative selection, the expectation over the clonotypes i of the autorecognition frequency is μ for all components k, and thus the expectation over the clonotypes of $\bar{w}_i$ is μ as well. One effect of negative selection is to introduce correlations between the elements of the law $\hat{\boldsymbol{L}}$ so that they can no longer be treated as independent. For instance, if it is known for some post-selection T cell with law $\hat{\boldsymbol{L}}$ that one of the elements of $\hat{\boldsymbol{L}}$ is larger than μ, there will be a reduced probability, relative to the unconditional value, that another element also exceeds μ.

Antigen Presentation Fluctuations

Differences between clonotypes are expressed by differences between the corresponding clonotype laws $\hat{\boldsymbol{L}}_i$. Their distribution reflects the statistical structure of the repertoire, and it is this distribution that negative selection acts upon. Within a given clonotype, statistical fluctuations remain; these are due to differences in APPs, which cause the TCR signal $w_{i\boldsymbol{\zeta}}$ to fluctuate about the clonotype mean $\bar{w}_i$. To characterise the within-clonotype fluctuations of the TCR signal across APPs, observe that component k contains N_k self pMHC species, out of which a number $n_{k\boldsymbol{\zeta}} \stackrel{\text{def}}{=} \sum_{j\in\mathscr{K}_k} \boldsymbol{I}_{j\boldsymbol{\zeta}}$ have a non-zero presentation level in APP $\boldsymbol{\zeta}$. Out of these $n_{k\boldsymbol{\zeta}}$ presented pMHC species, a number $\hat{n}_{ik\boldsymbol{\zeta}} = \sum_{j\in\mathscr{K}_k} \boldsymbol{I}_{j\boldsymbol{\zeta}} w_{ij}$ is recognised by a TCR of clonotype i. A standard argument shows that $\hat{n}_{ik\boldsymbol{\zeta}}$ follows the hypergeometric distribution:

$$\mathbb{P}(\hat{n}_{ik\boldsymbol{\zeta}} = x) = \frac{\binom{\hat{N}_{ik}}{x}\binom{N_k - \hat{N}_{ik}}{n_{k\boldsymbol{\zeta}} - x}}{\binom{N_k}{n_{k\boldsymbol{\zeta}}}} \tag{10}$$

on the assumption that peptides are independently presented. The description of the TCR signal $w_{i\boldsymbol{\zeta}}$ in the limit $N \to \infty$, Eq. (6), requires that the random variable $\hat{n}_{ik\boldsymbol{\zeta}}/N$ remains non-degenerate in this limit; this is shown in Appendix B.2, where a large deviations rate function for $\hat{n}_{ik\boldsymbol{\zeta}}/N_k$ is derived.

STATISTICS OF AUTOREACTIVITY FOLLOWING NEGATIVE SELECTION

This section presents three main results. The first result is a characterisation of the post-selection statistics of clonotype laws over the repertoire, in terms of a large deviations rate function. Then follows a second-order approximation of this rate function for an important class of clonotype statistics. The third result relates thymic presentation statistics to the role of negative selection in improving immune efficacy.

The Post-selection Large Deviations Rate Function

Let P_{surv} denote the probability that a thymocyte, chosen at random from among the thymocytes restricted to a given single MHC isoform, will appear in the mature naïve repertoire. This probability P_{surv} is not to be confused with the overall probability that a thymocyte will be allowed to mature. The latter includes both positive and negative selection, and its magnitude may largely reflect positive selection (Surh and Sprent, 1994; Merkenschlager *et al.*, 1997; Laufer *et al.*, 1999; Sebza *et al.*, 1999). In other words, P_{surv} denotes a thymocyte's probability of surviving into the mature repertoire conditional on being positively selected; it is given by

$$P_{\text{surv}} = \sum_{\boldsymbol{\nu}\in\mathscr{L}_N} \mathbb{P}\{\text{survival}|\hat{\boldsymbol{L}} = \boldsymbol{\nu}\}\mathbb{P}\{\hat{\boldsymbol{L}} = \boldsymbol{\nu}|\text{pre-selection}\}. \tag{11}$$

In this formula and those that follow, the term "selection" refers to negative selection only. The post-selection structure of the TCR repertoire can be discussed in terms of the probability that a mature T cell that has survived negative selection has a given law $\hat{\boldsymbol{L}}_i$. Bayes' rule relates this probability to pre-selection probabilities:

$$\mathbb{P}\{\hat{\boldsymbol{L}} = \boldsymbol{\nu}|\text{post-selection}\} =$$

$$\mathbb{P}\{\hat{\boldsymbol{L}} = \boldsymbol{\nu}|\text{pre-selection}\}\mathbb{P}\{\text{survival}|\hat{\boldsymbol{L}} = \boldsymbol{\nu}\}/P_{\text{surv}} \tag{12}$$

where $\boldsymbol{\nu} \in \mathscr{L}_N$ and i is the T cell's clonotype. The conditional survival probability is given by

$$\mathbb{P}\{\text{survival}|\hat{\boldsymbol{L}} = \boldsymbol{\nu}\} = \left[\mathbb{P}\{w_{i\boldsymbol{\zeta}} < w_{\text{thy}}|\hat{\boldsymbol{L}} = \boldsymbol{\nu}\}\right]^m \tag{13}$$

where w_{thy} is the thymic selection threshold of the thymocyte at hand and m denotes the number of antigen presentation events during which the T cell will undergo deletion if it registers a TCR signal greater than w_{thy} (Savage and Davis, 2001). Thus, m is the number of rounds of negative selection. Experimental estimates of m are not available, although there are some indications that thymocytes remain sufficiently long in the thymus to allow for dozens or perhaps even hundreds of rounds of negative selection (Scollay and Godfrey, 1995); this should however be set against the possibility that m is only of order 1, as may be the case for positive selection (Merkenschlager *et al.*, 1994; Merkenschlager, 1996).

Given a post-selection T cell which has successfully undergone m rounds of negative selection, the probability that this T cell's law will belong to an arbitrary open set $\Gamma \subset [0,1]^K$ can be estimated from the following large deviations result (Dembo and Zeitouni, 1998):

$$\lim_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\{\hat{\boldsymbol{L}} \in \Gamma|\text{post-selection}\} \stackrel{\text{def}}{=} I_\Gamma =$$

$$-\inf_{\boldsymbol{\nu}\in\Gamma}\left[H(\boldsymbol{\nu}|\boldsymbol{\mu}) + m \inf_{\boldsymbol{w}:\langle\mathbf{1},\boldsymbol{w}\rangle\le w_{\text{thy}}} \{\langle\boldsymbol{\pi}, \boldsymbol{I}(\boldsymbol{w};\boldsymbol{\nu},\boldsymbol{u},\boldsymbol{\rho})\rangle\} - I_{\text{surv}}\right] \tag{14}$$

where I_Γ is the post-selection rate function associated with the set Γ. Equation (14) shows how to find this rate function from the rate functions associated with fluctuations over clonotypes and over APPs: $H(\boldsymbol{\nu}|\mu)$ denotes the relative entropy of $\boldsymbol{\nu}$, defined in appendix B.1; $\boldsymbol{I}(\boldsymbol{w};\boldsymbol{\nu},\boldsymbol{u},\boldsymbol{\rho})$ is a K-vector of large deviations rate functions for the various components, defined in appendix B.2, and

$$I_{\text{surv}} \stackrel{\text{def}}{=} \lim_{N\to\infty} -\ln\{P_{\text{surv}}\}/N. \tag{15}$$

The evaluation of the infimum in Eq. (14) is somewhat involved, and is discussed in detail in appendix B.3.

Approximate Statistics of the Post-selection Repertoire

All the information regarding the post-selection TCR repertoire is contained in the large deviations rate function I_Γ. To understand the behaviour of this rate function one can consider its Taylor series about μ. This section gives leading-order estimates which afford insight into the relationship between the selection parameters and the presentation statistics of the negatively selecting cells.

The Survival Probability

The stringency of negative selection can be expressed as the probability P_{surv} that a thymocyte, chosen at random from among those thymocytes that are restricted to one MHC isoform, will appear in the mature naïve repertoire. One intuitively expects this survival probability P_{surv} to decrease with an increasing number of rounds of selection m and to increase with the selection threshold w_{thy}. This is borne out by the following estimates:

$$P_{\text{surv}} \approx \begin{cases} 1 & w_{\text{thy}} > \mu \\ \exp\left\{-N_{\boldsymbol{\rho}} \frac{m}{m+m_{\text{crit}}} \frac{(w_{\text{thy}}-\mu)^2}{2\mu}\right\} & w_{\text{thy}} < \mu \end{cases} \tag{16}$$

where $N_{\boldsymbol{\rho}} \stackrel{\text{def}}{=} N(\sum_{k=1}^K \rho_k^2/\pi_k)^{-1}$ and m_{crit} denotes the *critical number of presentation rounds,* itself defined in terms of the *effective mean ubiquity* $\bar{u}$:

$$m_{\text{crit}} \stackrel{\text{def}}{=} \frac{1-\bar{u}}{\bar{u}} \quad \text{where } \bar{u} \stackrel{\text{def}}{=} \frac{\sum_{k=1}^K \rho_k^2/\pi_k}{\sum_{k=1}^K \rho_k^2/(\pi_k u_k)}. \tag{17}$$

The effective mean ubiquity represents an average pMHC ubiquity in which presentation levels are taken into account, as can be seen more clearly from the following approximate formula in terms of data describing a representative APP $\boldsymbol{\zeta}$:

$$\bar{u} \approx \frac{\sum_{j=1}^N \zeta_j^2 u_j}{\sum_{j=1}^N \zeta_j^2}.$$

This shows how $\bar{u}$ may be determined empirically from antigen presentation data on the negatively selecting cells. Clearly, $\bar{u}$ is near 1 when most of the dominant, high $\bar{\zeta}$, self pMHC species are present in almost every APP of the negatively selecting cells, while $\bar{u}$ is almost zero when all pMHCs have a low probability of being presented on any given negatively selecting cell.

The effect of the number of presentation rounds depends on the typical ubiquity of the self pMHC species. In particular, when all pMHCs are always present ($\bar{u} = 1$), the across-APP variations vanish and a given thymocyte will register the same TCR signal on every conjugation, which means that $m_{\text{crit}} = 0$ and the value of m is immaterial as long as it is at least 1. On the other hand, when the typical ubiquity is very small, m_{crit} will be very large, and the survival probability will continue to decrease with m (for $m \lesssim m_{\text{crit}}$).

The number $N_{\boldsymbol{\rho}}$ can be determined from experimental data by means of the approximation

$$N_{\boldsymbol{\rho}} \approx n_{\text{div}} \big/ \bar{u}.$$

Here n_{div} denotes the presentation diversity, defined for a particular APP $\boldsymbol{\zeta}$ as follows:

$$n_{\text{div}} = 1 \Big/ \sum_{j=1}^N \zeta_j^2. \tag{18}$$

This presentation diversity is at least 1, a lower bound which is attained when a single pMHC species j' dominates with presentation level $\zeta_{j'} = 1$. Furthermore, n_{div} is at most the number of distinct pMHC species present in an APP, which is $N\langle\boldsymbol{\pi},\boldsymbol{u}\rangle \le N$; this upper bound is attained when all pMHC species present have the same presentation level.

The Clonotype Law after Negative Selection

The statistical distribution of clonotype laws among the post-selection T cells that make up the mature repertoire is of fundamental importance to immune response efficacy, tolerance and autoimmunity. This distribution can be characterised in terms of the distributions of a family of clonotype statistics of the form $\langle\boldsymbol{\alpha},\hat{\boldsymbol{L}}\rangle$ where $\boldsymbol{\alpha}$ lies in the simplex $\mathbb{S}$ defined by

$$\mathbb{S} \stackrel{\text{def}}{=} \left\{\boldsymbol{x} \in [0,1]^K : \sum_{k=1}^K x_k = 1\right\}. \tag{19}$$

Such a weighing vector $\boldsymbol{\alpha}$ can be used to represent the target of negative selection. This motivates the study of the across-repertoire distribution of statistics of the form $\langle\boldsymbol{\alpha},\hat{\boldsymbol{L}}\rangle$ before and after negative selection. In particular, if a vector $\tilde{\boldsymbol{\alpha}} \in \mathbb{S}$ can be found such that clonotypes with a high value of $\langle\tilde{\boldsymbol{\alpha}},\hat{\boldsymbol{L}}\rangle$ ought to be preferentially removed from the repertoire, the across-repertoire distribution of this statistic $\langle\tilde{\boldsymbol{\alpha}},\hat{\boldsymbol{L}}\rangle$ should be sharply truncated by the action of thymic selection. In this sense, $\tilde{\boldsymbol{\alpha}}$ represents the target of negative selection. Claim 1 below relates this target to thymic presentation of autoantigens.

Clonotype statistics of the form $\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle$ generalise the clonotype mean $\bar{w}_i$ (since $\bar{w}_i = \langle \boldsymbol{\rho}, \hat{\boldsymbol{L}}_i \rangle$ and $\boldsymbol{\rho} \in \mathbb{S}$). The post-selection statistical structure of the TCR repertoire may be studied by considering the family of distributions of the statistic $\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle$ parametrised by $\boldsymbol{\alpha} \in \mathbb{S}$, where $\hat{\boldsymbol{L}}$ is the law of a T cell chosen at random from the mature naïve repertoire. Approximations for this family of distributions follow shortly in Eqs. (22)–(24) below. These equations constitute a representation of the statistical structure of the repertoire after negative selection. The immunological interest is in exhibiting the roles played by the thymic presentation parameters and the selection parameters w_{thy} and m, and in establishing the principle of maximum selection efficiency (see "The efficiency of negative selection" section below).

Preliminary to the statement of the approximations, a few further compound parameters are defined. The following two coefficients relate the thymic presentation parameter $\boldsymbol{\rho}$ to the weighing vector $\boldsymbol{\alpha}$:

$$\varepsilon_{\boldsymbol{\alpha}} \stackrel{\text{def}}{=} \frac{\left(\sum_{k=1}^{K} \alpha_k \rho_k / \pi_k\right)^2}{\left(\sum_{k=1}^{K} \alpha_k^2 / \pi_k\right)\left(\sum_{k=1}^{K} \rho_k^2 / \pi_k\right)} \quad \text{and}$$

$$r_{\boldsymbol{\alpha}} \stackrel{\text{def}}{=} \frac{\sum_{k=1}^{K} \alpha_k^2 / \pi_k}{\sum_{k=1}^{K} \rho_k^2 / \pi_k}. \tag{20}$$

Finally, let $N_{\boldsymbol{\alpha}} \stackrel{\text{def}}{=} N(\sum_{k=1}^{K} \alpha_k^2 / \pi_k)^{-1} = N_{\boldsymbol{\rho}} / r_{\boldsymbol{\alpha}}$ and let

$$\omega_{\text{crit}} \stackrel{\text{def}}{=} \begin{cases} \mu + \sqrt{r_{\boldsymbol{\alpha}} \varepsilon_{\boldsymbol{\alpha}}}(w_{\text{thy}} - \mu)/(1 + m_{\text{crit}}/m) & \text{for } w_{\text{thy}} \le \mu \\ \mu + \sqrt{r_{\boldsymbol{\alpha}} / \varepsilon_{\boldsymbol{\alpha}}}(w_{\text{thy}} - \mu) & \text{for } w_{\text{thy}} > \mu. \end{cases} \tag{21}$$

Large deviations theory only describes the effect of $m \ge 0$ rounds of negative selection for $\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle > \omega_{\text{crit}}$. In the case where $w_{\text{thy}} \le \mu$, the following approximation obtains: for $\omega \le \omega_{\text{crit}}$, $\mathbb{P}\{\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle > \omega | \, m \text{ rounds}\} \approx 1$ whereas for $\omega > \omega_{\text{crit}}$,

$$\mathbb{P}\{\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle > \omega | m \,\text{rounds}\}$$

$$\approx \exp\left\{ \frac{-\mathrm{N}_{\boldsymbol{\alpha}}}{2\mu} \left((\omega - \mu)^2 + \frac{\left[\sqrt{\varepsilon_{\boldsymbol{\alpha}}}(\omega - \mu) - \sqrt{r_{\boldsymbol{\alpha}}}(w_{\text{thy}} - \mu)\right]^2}{1 - \varepsilon_{\boldsymbol{\alpha}} + m_{\text{crit}}/m} \right.\right.$$

$$\left.\left. \text{hskip38pt} - \frac{r_{\boldsymbol{\alpha}}(w_{\text{thy}} - \mu)^2}{1 + m_{\text{crit}}/m} \right) \right\}.$$

When $w_{\text{thy}} > \mu$, the approximation is slightly more complicated: for $\omega \le \mu$, $\mathbb{P}\{\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle > \omega \,|\, m \text{ rounds}\} \approx 1$; next, for $\mu < \omega \le \omega_{\text{crit}}$,

$$\mathbb{P}\{\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle > \omega \,|\, m \text{ rounds}\} \approx \exp\left\{ -N_{\boldsymbol{\alpha}} \frac{(\omega - \mu)^2}{2\mu} \right\} \tag{23}$$

and, finally, for $\omega > \omega_{\text{crit}}$,

$$\mathbb{P}\{\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle > \omega \,|\, m \text{ rounds}\}$$

$$\approx \exp\left\{ \frac{-N_{\boldsymbol{\alpha}}}{2\mu} \left((\omega - \mu)^2 + \frac{\left[\sqrt{\varepsilon_{\boldsymbol{\alpha}}}(\omega - \mu) - \sqrt{r_{\boldsymbol{\alpha}}}(w_{\text{thy}} - \mu)\right]^2}{1 - \varepsilon_{\boldsymbol{\alpha}} + m_{\text{crit}}/m} \right) \right\}. \tag{24}$$

The pre-selection approximations are obtained for $m = 0$ in these expressions: in that case, $\omega_{\text{crit}} = \mu$, and $\mathbb{P}\{\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle > \omega \,|\, m \text{ rounds}\} \approx 1$ for $\omega \le \mu$, while Eq. (23) holds for $\omega > \mu$.

Post-selection Autorecognition Frequencies

Prior to negative selection, the average autorecognition frequency $\hat{L}_k$ among the thymocytes is μ for every component k. Negative selection reduces this average by rendering high autorecognition frequencies less likely. The autorecognition frequency is a statistic of the form $\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle$ with $\boldsymbol{\alpha} = \mathbf{1}_k$, that is, $\alpha_k = 1$ and $\alpha_\ell = 0$ for $\ell \ne k$:

$$\hat{L}_k = \langle \mathbf{1}_k, \hat{\boldsymbol{L}} \rangle.$$

In the case where $w_{\text{thy}} > \mu$, large deviations in autorecognition frequencies after m rounds of negative selection are described by

$$\mathbb{P}\{\hat{L}_k > \omega \,|\, m \text{ rounds}\}$$

$$\approx \exp\left\{ -\frac{\pi_k N}{2\mu} \left((\omega - \mu)^2 + \frac{\varepsilon_{\mathbf{1}_k}(\omega - \omega_{\text{crit}})^2}{1 - \varepsilon_{\mathbf{1}_k} + m_{\text{crit}}/m} \right) \right\}$$

for $\omega > \omega_{\text{crit}}$, where

$$\omega_{\text{crit}} = \mu + \frac{w_{\text{thy}} - \mu}{\rho_k} \quad \text{and} \quad \varepsilon_{\mathbf{1}_k} = \frac{\rho_k^2 / \pi_k}{\sum_{\ell=1}^{K} \rho_\ell^2 / \pi_\ell}.$$

Together, these expressions indicate that a sharp truncation at $\hat{L}_k = w_{\text{thy}}$ can be achieved only if component k dominates thymic presentation, with ρ_k close to 1. Otherwise, the truncating effect is imperfect and becomes prominent only at higher values of the autorecognition frequency. If the component is not prominent at all during thymic selection, the post-selection distribution will hardly differ from the pre-selection distribution. For such a component, self-nonself discrimination fails, and its self pMHC species are effectively foreign. Indeed, thymic expression of autoantigens has been found to correlate with resistance to autoimmune disease (Egwuagu *et al.*, 1997). Similar conclusions apply in the case where $w_{\text{thy}} < \mu$. These observations are generalised in the next section.

The Efficiency of Negative Selection

The extent to which negative selection in the thymus has modified the TCR repertoire can be gauged by studying

the post-selection distribution of $\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle$: varying $\boldsymbol{\alpha}$ within the simplex $\mathbb{S}$, one obtains a family of distributions $\mathbb{P}\{\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle > \omega\}$ which probe the post-selection structure of the TCR repertoire. Negative selection reduces the probability that $\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle$ exceeds a given ω. The extent of the reduction depends on both $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$, and is maximised when $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ coincide, as is expressed by the following claim.

Claim 1 (Maximum Efficiency of Negative Selection) Given thymic presentation statistics $\boldsymbol{u}$, a fixed selection parameter w_{thy}, a fixed but sufficiently large number of self peptides N, a weighing vector $\boldsymbol{\alpha} \in \mathbb{S}$, and $\omega > w_{\text{thy}}$ with ω sufficiently close to μ, the probability

$$\mathbb{P}\{\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle > \omega | \text{ after } m \text{ rounds of negative selection}\}$$

can be made arbitrarily small by choosing m sufficiently large and the thymic presentation vector $\boldsymbol{\rho}$ sufficiently close to $\boldsymbol{\alpha}$.

On the approximations of Eqs. (22) and (24), the large deviations rate function can be made arbitrarily large only if the quantity

$$1 - \varepsilon_{\boldsymbol{\alpha}} + m_{\text{crit}}/m$$

is made arbitrarily small (this follows from the fact that both $r_{\boldsymbol{\alpha}}$ and $\varepsilon_{\boldsymbol{\alpha}}$ have finite upper and lower bounds as $\boldsymbol{\rho}$ varies over $\mathbb{S}$, with $\boldsymbol{\alpha}$ fixed). This is shown in Fig. 1. By Hölder's inequality,

$$\varepsilon_{\boldsymbol{\alpha}} \leq 1$$

with equality only when $\boldsymbol{\rho} = \boldsymbol{\alpha}$. Thus, $1 - \varepsilon_{\boldsymbol{\alpha}}$ is smaller than a given positive bound iff $\boldsymbol{\rho}$ is sufficiently close to $\boldsymbol{\alpha}$, and Claim 1 follows. The claim must be restricted to $\omega > w_{\text{thy}}$, since $\omega_{\text{crit}} \to w_{\text{thy}}$ when $\boldsymbol{\rho} \to \boldsymbol{\alpha}$. The present proof is based on a second-order approximation of a large deviations estimate, which imposes the following technical restrictions: (i) the number of self peptides N has to be large enough for the large deviations-based approximation to be relevant; and (ii) w_{thy} and ω have to be sufficiently close to μ to ensure that the large deviations function is dominated by its second-order term.

In general, the value of w_{thy} must be expected to vary among thymocytes, and it may even change for a given thymocyte as it matures. In view of these variations, it is significant that Claim 1 does not hinge on the value of w_{thy} but rather on the characteristics m and $\boldsymbol{\rho}$, which are governed by thymic architecture and differentiation of the thymic stroma (Laufer *et al.*, 1999; Sebza *et al.*, 1999).

According to Claim 1 the thymic presentation vector $\boldsymbol{\rho}$ encodes, effectively, the target of negative selection. The relationship between thymic presentation and immune efficacy only makes sense in the light of the statistical fluctuations of peripheral self presentation. This is explained in the next section by means of two illustrations of how Claim 1 can be applied.

IDENTIFYING THE TARGET OF NEGATIVE SELECTION

The efficiency result, Claim 1, implies that the thymic presentation vector $\boldsymbol{\rho}$ is an objective representation of the target of negative selection. In particular, among all clonal statistics of the form $\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle$ with $\boldsymbol{\alpha} \in \mathbb{S}$, the one that is most efficiently selected against is $\langle \boldsymbol{\rho}, \hat{\boldsymbol{L}} \rangle$. To understand the immunological significance of the thymic presentation vector, it is useful to apply the efficiency result in the opposite direction: that is, formulate a hypothesis concerning the target of negative selection, encode it as a weighing vector $\boldsymbol{\alpha} \in \mathbb{S}$, and compare this to the actual thymic presentation $\boldsymbol{\rho}$. While this is, in principle, a feasible empirical programme, difficulties may arise given experimental limitations in assessing presentation levels and ubiquities of a large sample of self peptides, among

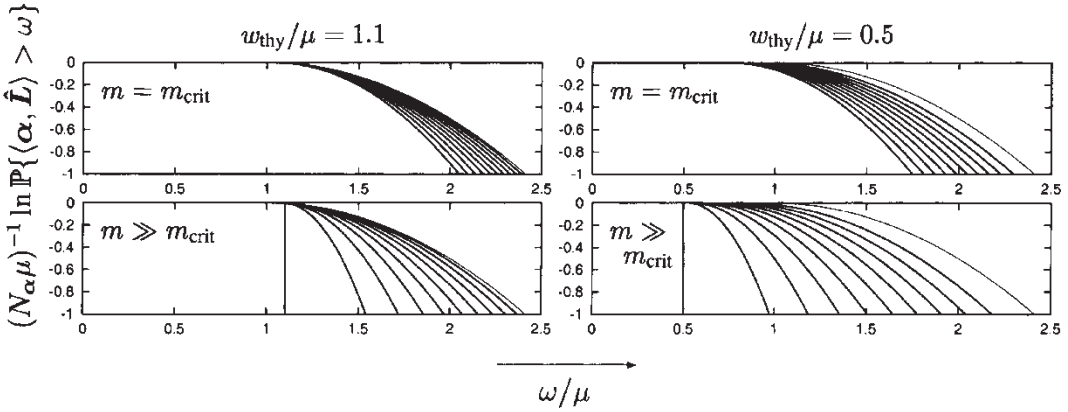

FIGURE 1 Efficiency of negative selection. The probability that the clonal statistic $\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle$ exceeds ω, plotted as a function of ω/μ. Dotted line indicates the pre-selection case. Curves are from left to right for decreasing $\varepsilon_{\boldsymbol{\alpha}}$, a parameter which expresses how well $\boldsymbol{\alpha}$ and the thymic presentation vector $\boldsymbol{\rho}$ are matched; $\varepsilon_{\boldsymbol{\alpha}} = 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1$. Top panels: the number of rounds of negative selection m equals $m_{\text{crit}} \stackrel{\text{def}}{=} (1 - \bar{u})/\bar{u}$ where $\bar{u}$ is the effective mean ubiquity of thymic presentation; bottom panels: limiting case where m is much larger than m_{crit}. Left panels: $w_{\text{thy}} = 1.1\,\mu$; right panels: $w_{\text{thy}} = 0.5\,\mu$. All panels: $r_{\boldsymbol{\alpha}} = 1$ $\left(N_{\boldsymbol{\alpha}} \stackrel{\text{def}}{=} N(\sum_{k=1}^{K} \alpha_k^2/\pi_k)^{-1}\right)$.

both the classes of negatively selecting cells and secondary lymphoid pAPCs.

The presentation statistics of the professional APCs (pAPCs) in the secondary lymphoid tissues are indicated with a prime: $\boldsymbol{\rho}', \boldsymbol{u}'$. Derived statistics such as presentation diversity n'_{div} and mean ubiquity $\bar{u}'$ are important because they can be determined from experimental data. Approximate formulæ are:

$$\bar{u}' \approx \frac{\sum_{j=1}^{N} \zeta'^2_j u'_j}{\sum_{j=1}^{N} \zeta'^2_j} \quad \text{and} \quad n'_{\text{div}} \approx \frac{1}{\sum_{j=1}^{N} \zeta'^2_j},$$

to be determined from data on a representative APP $\boldsymbol{\zeta}'$ from a secondary lymphoid pAPC.

In the present section, the programme of identifying the target of negative selection will be illustrated by an application to two hypotheses concerning the target of negative selection. Whereas the first of these two hypotheses (negative selecting is means-directed) probably reflects a consensus among immunologists, the second (negative selection is variance-directed) is perhaps more controversial.

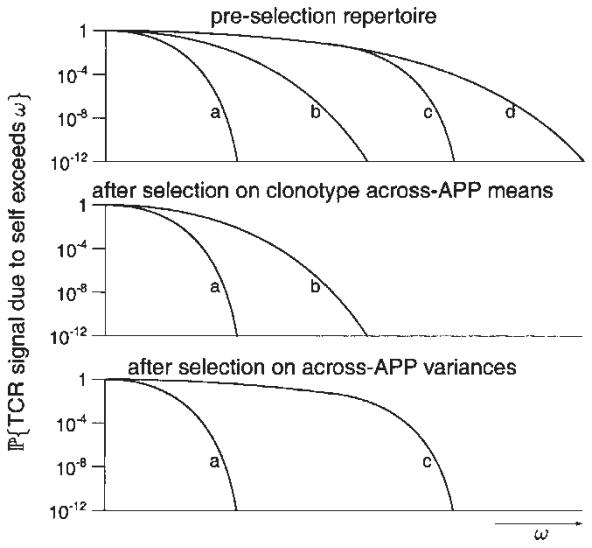

FIGURE 2 Two hypotheses concerning the target of negative selection. The probability that the TCR signal due to self exceeds ω as peripheral APCs vary, plotted as a function of ω for four representative clonotypes: (a) a clone with a low mean and small variance; (b) a clone with a low mean and large variance; (c) a clone with a large mean and small variance; (d) a clone with a large mean and large variance. Top panel: situation prior to negative selection. Middle panel: situation after selection targeted at clonotype means. Bottom panel: situation after selection targeted at clonotype variances.

Means- versus Variance-directed Negative Selection

Two natural parameters of a mature recirculating T cell's autoreactivity are the average and the variance of the TCR signal due to self registered by the T cell as it interacts with pAPCs in secondary lymphoid tissues throughout the body. While both parameters are *prima facie* a candidate target for central tolerance, the post-selection structure of the T cell repertoire would be markedly different, as shown schematically in Fig. 2.

Means-directed Negative Selection

A mature recirculating T cell with autorecognition frequencies $\hat{\boldsymbol{L}}$ experiences an average TCR signal $\langle \boldsymbol{\rho}', \hat{\boldsymbol{L}} \rangle$ from self peptides on the pAPCs it encounters. A natural assumption is that negative selection acts to remove the T cells with the highest means $\langle \boldsymbol{\rho}', \hat{\boldsymbol{L}} \rangle$. This assumption identifies the weighing vector $\boldsymbol{\alpha}$ with the peripheral presentation vector $\boldsymbol{\rho}'$. The maximum efficiency principle, Claim 1, then implies $\boldsymbol{\rho} = \boldsymbol{\rho}'$ for the thymic presentation vector. The approximations presented in the previous section then take on a much simpler form: for $w_{\text{thy}} \leq \mu$ and $\omega > \omega_{\text{crit}} = \mu + (w_{\text{thy}} - \mu)m/(m + m_{\text{crit}})$, Eq. (22) becomes:

$$\mathbb{P}\{\langle \boldsymbol{\rho}', \hat{\boldsymbol{L}} \rangle > \omega \,|\, m \text{ rounds}\}$$
$$\approx \exp\left\{ -\frac{n'_{\text{div}}/\bar{u}'}{2\mu} \left((\omega - \mu)^2 + \frac{m}{m_{\text{crit}}} (\omega - w_{\text{thy}})^2 - \frac{(w_{\text{thy}} - \mu)^2}{1 + m_{\text{crit}}/m} \right) \right\}, \quad (25)$$

whereas for $w_{\text{thy}} > \mu$ and $\omega > w_{\text{thy}}$, Eq. (24) becomes:

$$\mathbb{P}\{\langle \boldsymbol{\rho}', \hat{\boldsymbol{L}} \rangle > \omega | m \text{ rounds}\}$$
$$\approx \exp\left\{ -\frac{n'_{\text{div}}/\bar{u}'}{2\mu} \left((\omega - \mu)^2 + \frac{m}{m_{\text{crit}}} (\omega - w_{\text{thy}})^2 \right) \right\}. \quad (26)$$

These expressions clearly show that the number of presentation rounds m must be of order m_{crit} or larger to obtain a sizable post-selection improvement of the TCR repertoire, as is shown in Fig. 3. The critical number of presentation rounds m_{crit} decreases with increasing thymic effective mean ubiquity $\bar{u}$. However, $\boldsymbol{\rho} = \boldsymbol{\rho}'$ implies that $n_{\text{div}}/\bar{u} = n'_{\text{div}}/\bar{u}'$. Thus, if the mean ubiquity in the thymus $\bar{u}$ is higher than the peripheral value $\bar{u}'$, this must be compensated for by a proportionally greater presentation diversity n_{div} on the negatively selecting cells. This, in turn, would require high MHC counts on the negatively selecting cells to safeguard signalling fidelity (see van den Berg *et al.*, 2001 and van den Berg and Rand, 2003 for detailed arguments).

Variance-directed Negative Selection

An alternative hypothesis is that negative selection is directed against the variance of $w_{i\boldsymbol{\zeta}}$, the TCR signal across the APPs on professional APCs in the secondary lymphoid tissues. Thus, thymic selection particularly targets those clonotypes whose across-APP variance in the periphery is large. The question then becomes how thymic

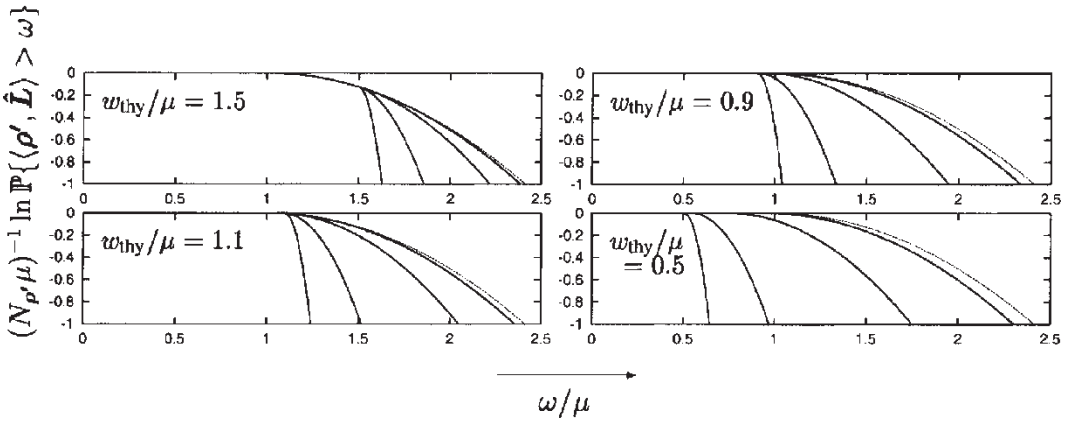

FIGURE 3 Means-directed negative selection. The probability that the mean peripheral TCR signal due to self $\langle \boldsymbol{\rho}', \hat{\boldsymbol{L}} \rangle$ exceeds ω, plotted as a function of ω/μ. Dotted line indicates the pre-selection case. Curves are from left to right for decreasing values of $m/m_{\rm crit}$, the number of rounds of negative selection, m, relative to a critical number $m_{\rm crit} \stackrel{\rm def}{=} (1-\bar{u})/\bar{u}$ where $\bar{u}$ is the effective mean ubiquity of thymic presentation; $m/m_{\rm crit} = 100, 10, 1, 0.1$. Top left: $w_{\rm thy} = 1.5\,\mu$; Bottom left: $w_{\rm thy} = 1.1\,\mu$; Top right: $w_{\rm thy} = 0.9\,\mu$; Bottom right: $w_{\rm thy} = 0.5\,\mu$. ($N_{\boldsymbol{\rho}'} \stackrel{\rm def}{=} N\big(\sum_{k=1}^{K} \rho'_k 2/\pi_k\big)^{-1}$).

presentation statistics must be skewed relative to those of the periphery in order to effect variance-directed selection most efficiently; this question is answered by the "skew formula" given in Eq. (28) below. If experimental results on thymic presentation correspond to this formula, this would count as evidence in support of variance-directed selection.

The variance ${\sigma'}^2_{\rm APP}$ is well-approximated for large N by

$$ {\sigma'}^2_{\rm APP} \approx \frac{1}{N} \sum_{k=1}^{K} \frac{{\rho'}_k^2 (1-u'_k)}{\pi_k u'_k} \hat{L}_k (1-\hat{L}_k). \qquad (27) $$

Since $\hat{L}_k \ll 1$ for all components k, the across-APP variance (given by Eq. (27) above) can be approximated by a linear combination of the autorecognition frequencies:

$$ {\sigma'}^2_{\rm APP} \approx \frac{1}{N} \sum_{k=1}^{K} \frac{{\rho'}_k^2 (1-u'_k)}{\pi_k u'_k} \hat{L}_k. $$

The maximum efficiency principle $\boldsymbol{\rho} = \boldsymbol{\alpha}$ implies that ρ_k should be proportional to the coefficient of $\hat{L}_k$ in the above sum, which yields the following expression for each pair of components k and ℓ:

$$ \frac{\rho_k}{\rho_\ell} = \frac{{\rho'}_k^2 (1-u'_k)/(\pi_k u'_k)}{{\rho'}_\ell^2 (1-u'_\ell)/(\pi_\ell u'_\ell)} = \frac{\rho'_k (1-u'_k) \bar{\zeta}'_k}{\rho'_\ell (1-u'_\ell) \bar{\zeta}'_\ell}, \qquad (28) $$

which shows that the thymic presentation $\boldsymbol{\rho}$ is similar to the peripheral presentation $\boldsymbol{\rho}'$, but skewed in favour of components with high peripheral pAPC peptide presentation levels $\bar{\zeta}'_k$ as well as components with low ubiquities $\bar{u}'_k$ in these pAPCs. In particular, it is possible that component k dominates ℓ in secondary lymphoid pAPCs ($\rho'_k > \rho'_\ell$) while the situation is reversed in the thymus ($\rho_k < \rho_\ell$) as a result of component k having peripherally a lower ubiquity or higher mean presentation level, or both. The prediction, Eq. (28), can be rewritten in terms of relative pMHC presentation levels instead of component presentation levels:

$$ \frac{\bar{\zeta}_k}{\bar{\zeta}_\ell} = \frac{u'_k (1-u'_k)/u_k}{u'_\ell (1-u'_\ell)/u_\ell} \left(\frac{\bar{\zeta}'_k}{\bar{\zeta}'_\ell} \right)^2 \qquad (29) $$

thus if the thymic ubiquities are the same, thymic presentation levels will be high for components with intermediate peripheral ubiquities (near 1/2) and/or high relative presentation levels when expressed in the periphery.

Two lines of evidence argue in favour of the hypothesis that negative selection is directed against across-APP variance ${\sigma'}^2_{\rm APP}$, rather than the mean $\langle \boldsymbol{\rho}', \hat{\boldsymbol{L}} \rangle$. The first is that, at least on a qualitative level, the available evidence on thymic presentation is consistent with the following prediction, derived from Eq. (28): thymic presentation is skewed in favour of self proteins which are (i) expressed only in specific tissues, but where expressed, (ii) expressed in large quantities. Property (i) entails a low ubiquity, since the statistics need to be defined over all the body's secondary lymphoid tissues (this follows from the random recirculation dynamics of a naïve T cell), whereas property (ii) promotes large relative presentation levels per peptide, *ceteris paribus*. Protein hormones constitute a good example: ectopic thymic presentation of such hormones has been reported in the literature (Hanahan, 1998; Sospreda *et al.*, 1998).

The theory furnishes a direct way to confirm the hypothesis that negative selection is directed against across-APP variance: Eqs. (28) and (29) predict which self peptides can be expected to be over-represented in the thymic presentation patterns, and by how much. A detailed experimental verification would require the determination of antigen presentation patterns and ubiquities in two classes of APCs (negatively selecting cells in the thymus and pAPCs in the secondary lymphoid tissues). Although perhaps arduous and costly, this is not an impossible task. Antigen presentation patterns can be determined (e.g. Hunt *et al.*, 1992) and ubiquities

can be estimated by identifying peptides on pAPCs, notably on APCs taken from lymph nodes draining a variety of tissues. Of course, accurate determination of a low ubiquity may be tedious, as is true in general of the empirical determination of low frequencies.

A second line of evidence in favour of variance-directed selection is theoretical. It rests on the following claim:

CLAIM 2 (VARIANCE-DIRECTED NEGATIVE SELECTION) Negative selection directed against the across-APP peripheral variance σ'^2_{APP} maximises the sensitivity of the TCR repertoire to foreign antigen while minimising the risks of (i) mounting an immune response against self peptides and (ii) failing to mount an immune response against the foreign peptide.

The following section makes the terminology of this claim more precise, and offers an argument to support it.

Variance-directed Selection and Immune Response Efficacy

Central to both the problems of autoimmunity and response efficacy is the probability that a mature T cell is activated by a pAPC presenting a foreign peptide while the T cell does not actually recognise the foreign peptide:

$$\mathbb{P}\{w_{i\tilde{\jmath}} = 0 | \text{clone } i \text{ has been activated}\}$$

where $\tilde{\jmath}$ is the foreign peptide. This conditional probability is the expected fraction of clones among those initially activated which do not recognise the foreign peptide and are potentially harmful.

When the foreign (non-self) peptide $\tilde{\jmath}$ occupies a fraction $\zeta_{\tilde{\jmath}}$ of the MHC molecules capable of interacting with the TCR at hand, the TCR signal becomes:

$$w_{i\boldsymbol{\zeta}} = \zeta_{\tilde{\jmath}} w_{i\tilde{\jmath}} + (1 - \zeta_{\tilde{\jmath}}) \sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{ik\boldsymbol{\zeta}} \tag{30}$$

on the assumption that all components are "displaced" equally by the foreign peptide (cf. van den Berg *et al.*, 2001). From Eq. (30) one can derive

$$\mathbb{P}\{w_{i\tilde{\jmath}} = 0 | \text{clone } i \text{ has been activated}\} = \frac{\mathbb{P}\{\sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{ik\boldsymbol{\zeta}} > w_{\mathrm{act}}/(1 - \zeta_{\tilde{\jmath}})\}}{\mathbb{P}\{\sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{\mathrm{ik}\boldsymbol{\zeta}} > w_{\mathrm{act}}/(1 - \zeta_{\tilde{\jmath}})\} + \frac{\mu}{1-\mu} \mathbb{P}\{\sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{ik\boldsymbol{\zeta}} > (w_{\mathrm{act}} - \zeta_{\tilde{\jmath}})/(1 - \zeta_{\tilde{\jmath}})\}}.$$

The fraction on the right is very small if

$$\frac{\mu}{1-\mu} \mathbb{P}\left\{\sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{ik\boldsymbol{\zeta}} > \frac{w_{\mathrm{act}} - \zeta_{\tilde{\jmath}}}{1 - \zeta_{\tilde{\jmath}}}\right\} \gg \mathbb{P}\left\{\sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{ik\boldsymbol{\zeta}} > \frac{w_{\mathrm{act}}}{1 - \zeta_{\tilde{\jmath}}}\right\}$$

that is, the probabilities of the self background exceeding $w_{\mathrm{act}} - \zeta_{\tilde{\jmath}}$ and w_{act} are to be well-separated:

$$\mathbb{P}\left\{\sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{ik\boldsymbol{\zeta}} > w_{\mathrm{act}} - \zeta_{\tilde{\jmath}}\right\} \gg \mu^{-1} \mathbb{P}\left\{\sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{ik\boldsymbol{\zeta}} > w_{\mathrm{act}}\right\} \tag{31}$$

(with $1 - \zeta_{\tilde{\jmath}} \approx 1$ and $\mu/(1-\mu) \approx \mu$). This separation decreases as the within-clonotype variance increases. Condition (31) shows that the separation between the probabilities is smaller at low foreign presentation levels $\zeta_{\tilde{\jmath}}$. This effect puts a lower limit on the foreign presentation levels that can be safely detected. Thus, a T cell with a small across-APP variance σ'^2_{APP} achieves a greater separation in the sense of condition (31) at a given foreign presentation level $\zeta_{\tilde{\jmath}}$. It follows that *the immune system as a whole will be better able to detect foreign antigen at low levels if the repertoire is selected for low across-APP variances in the secondary lymphoid tissues*.

The minimum foreign presentation level that can be reliably and safely detected by a given T cell is in fact proportional to σ'_{APP} for that T cell. This can be derived by combining the requirements that (i) autoimmunity be avoided and (ii) at least one clonotype is activated with high probability. The details of this derivation are as follows:

The *autoimmunity probability* is the probability that a T cell is activated by a pAPC presenting autoantigens only (in addition to appropriate signal 2 co-stimulation). The requirement of non-responsiveness to self can be formulated in terms of an upper bound on this autoimmunity probability:

$$\mathbb{P}\left\{\sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{ik\boldsymbol{\zeta}} > w_{\mathrm{act}}\right\} \leq \boldsymbol{\epsilon}, \tag{32}$$

where $\boldsymbol{\epsilon}$ denotes a maximum allowable risk of autoimmunity, which induces an associated tolerance threshold $w_{\boldsymbol{\epsilon}}$ through

$$\mathbb{P}\left\{\sum_{k=1}^{K} \bar{\zeta}'_k \hat{n}_{ik\boldsymbol{\zeta}} > w_{\boldsymbol{\epsilon}}\right\} = \boldsymbol{\epsilon}. \tag{33}$$

Consequently, autoimmunity is avoided up to risk level $\boldsymbol{\epsilon}$ if the T cell's activation threshold satisfies

$$w_{\mathrm{act}} \geq w_{\boldsymbol{\epsilon}}.$$

Now consider a number N_T of naïve T cells, all of distinct clonotype, and which all have the same activation threshold w_{act}. The expected number of responders among

these N_T cells is found by multiplying N_T with the activation probability $P_{\mathrm{act}}(w_{\mathrm{act}};\zeta_{\tilde{\jmath}})$, where

$$P_{\mathrm{act}}(\omega;\zeta_{\tilde{\jmath}}) \stackrel{\mathrm{def}}{=} \mathbb{P}\left\{\zeta_{\tilde{\jmath}} w_{i\tilde{\jmath}} + (1-\zeta_{\tilde{\jmath}})\sum_{k=1}^{K}\bar{\zeta}'_k\hat{n}_{ik\zeta} > \omega\right\}. \tag{34}$$

A graph of $P_{\mathrm{act}}(\omega;\zeta_{\tilde{\jmath}})$ against ω for a fixed value of $\zeta_{\tilde{\jmath}}$ exhibits a plateau at $P_{\mathrm{act}} \approx \mu$ of width $\sim \zeta_{\tilde{\jmath}}$, as shown in Fig. 4. The existence of this plateau means that the immune system can robustly operate at a typical activation probability of μ, provided that the T cells in the repertoire all have cellular activation values within the range of this plateau.

To satisfy the requirement that almost certainly at least one clone responds, there must be well over $1/\mu$ distinct clonotypes in the repertoire for each MHC isoform; the probability of failing to activate any one of N_T T cells of distinct clonotypes is $\exp\{-\mu N_T\}$, since the number of responding clonotypes is Poisson distributed with mean μN_T.

Finally, observe from the construction in Fig. 4 that the requirements $P_{\mathrm{act}}(w_{\mathrm{act}};\zeta_{\tilde{\jmath}}) \geq \mu$ and $w_{\mathrm{act}} \geq w_\epsilon$ can only be simultaneously fulfilled if

$$\zeta_{\tilde{\jmath}} \geq \zeta_{\min} \quad \text{with} \quad \zeta_{\min} \stackrel{\mathrm{def}}{=} w_\epsilon - w_\mu$$

where w_μ is implicitly defined by

$$\mathbb{P}\left\{\sum_{k=1}^{K}\bar{\zeta}'_k\hat{n}_{ik\zeta} > w_\mu\right\} = \mu. \tag{35}$$

Thus $\zeta_{\min} = w_\epsilon - w_\mu$ represents the lowest foreign presentation level which can be safely and robustly detected. Combining Eqs. (33) and (35) with the leading-order estimate

$$\ln \mathbb{P}\left\{\sum_{k=1}^{K}\bar{\zeta}'_k\hat{n}_{ik\zeta} > \omega\right\} \approx -\frac{(\langle \boldsymbol{\rho}', \hat{\boldsymbol{L}}\rangle - \omega)^2}{2\sigma'^2_{\mathrm{APP}}}$$

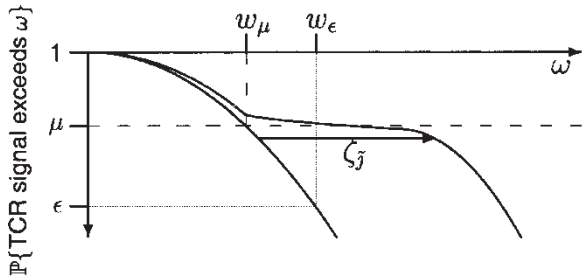

FIGURE 4 Diagram to derive the minimum safely detectable foreign presentation level. The probability that the TCR signal exceeds ω, plotted as a function of ω. Left curve: TCR signal due to self alone; right curve: TCR signal due to self plus foreign peptide presented at a relative presentation level of $\zeta_{\tilde{\jmath}}$. For a T cell with cellular activation threshold between w_μ and $w_\mu + \zeta_{\tilde{\jmath}}$ the probability of activation is μ, the probability that its TCR recognises the foreign epitope. Avoidance of autoimmunity requires that the cellular activation threshold exceeds w_ϵ: this leads to the requirement that $\zeta_{\tilde{\jmath}}$ be at least $w_\epsilon - w_\mu$. See text for further explanation.

(see van den Berg and Rand, 2003 for a derivation) one finds the following estimate for the minimum detectable foreign presentation level:

$$\zeta_{\min} \approx \kappa\sigma'_{\mathrm{APP}} \quad \text{where}$$
$$\kappa = \sqrt{2\ln\{1/\epsilon\}} - \sqrt{2\ln\{1/\mu\}}. \tag{36}$$

Thus T cells with lower across-APP variances can detect lower foreign presentation levels, given a certain autoimmunity risk level ϵ.

Variance-directed negative selection is not necessarily at odds with the consensus view that abundantly expressed autoantigens are targeted by negative selection. The idea is rather that thymic presentation should be skewed towards the less abundant (i.e. low-ubiquity) autoantigens. Moreover, thymic presentation should, according to Eq. (29), also be skewed towards the autoantigens that are presented at the highest levels on the pAPCs found in the secondary lymphoid tissues.

A paradoxical property of variance-directed selection is that pMHC species that occur constitutively on all secondary lymphoid pAPCs (that is, with ubiquity 1) are not presented at all in the thymus, as is shown by Eq. (28): if component k is constitutive, $u_k{}' = 1$, it follows that $\rho_k = 0$. Of course it may not be practically possible for the negatively selecting cells to completely suppress presentation of constitutive peptides that belong to household proteins. However, suppose for the sake of the argument that negatively selecting cells are able to put variance-directed selection into effect exactly as prescribed by Eq. (28), and consider a hypothetical T cell which recognises only constitutive autoantigens, possibly many of them. Such a T cell does not encounter any of its recognised autoantigens during negative selection, even though it is highly autoreactive. The present argument resolves this paradox by insisting that it is across-APP variance that hampers T cell efficacy, and that constitutive pMHCs do not contribute to this variance. Nevertheless, it is essential that the activation threshold w_{act} of the T cell be higher than $\bar{w}_i$. This indicates that variance-directed central tolerance must be complemented by additional tolerance mechanisms to ensure that the threshold satisfies $w_{\mathrm{act}} > w_\epsilon \geq \bar{w}_i$.

Central versus Peripheral Tolerance

The hypothesis of variance-directed negative selection accords a crucial role to peripheral mechanisms of tolerance: to ensure that $w_{\mathrm{act}} > w_\epsilon$ for all T cells in the mature repertoire. An active mechanism is required since every clonotype has its own value of w_ϵ, as can be seen from the probability in Eq. (33) which defines w_ϵ and depends on the clonotype law governing the distribution of $\hat{n}_{ik\zeta}$. There will be a natural variation in w_{act} values among the T cells that leave the thymus, and associated with each value is an autoimmunity probability

$\mathbb{P}\{\sum_{k=1}^{K}\bar{\zeta}_k\hat{n}_{ik\zeta}\} > w_{\mathrm{act}}$. This probability is reduced by the action of peripheral tolerance mechanisms.

Broadly speaking, two modes of action for peripheral tolerance can be envisaged, both of which are equally compatible with variance-directed selection. Both act by evoking a T cell response to a suprathreshold TCR signal in the absence of an appropriate immunostimulatory environment (Webb *et al.*, 1990; Janeway and Travers, 1997; Gallucci and Matzinger, 2001; Hawiger *et al.*, 2001). In the selective mode, the T cell is induced to enter apoptosis (Webb *et al.*, 1990; Hawiger *et al.*, 2001), whereas in the instructive mode, the T cell is induced to adapt its w_{act} value upward (Wong *et al.*, 2001), resulting in a diminished responsiveness or, ultimately, anergy, which may be an intermediate stage on the way to the suppressor phenotype (Webb *et al.*, 1990; Grossman and Paul, 1992; Lombardi *et al.*, 1994; Roncarolo and Levings, 2000). For the present purposes it is not necessary to assess the relative importance of the two modes, since both are compatible with the variance-hypothesis. The instructive mode is inherently less wasteful, and ties in better with the evidence in support of dynamic, tunable cellular activation thresholds (Grossman and Singer, 1996; Nicholson *et al.*, 2000; Grossman and Paul, 2001).

Both modes of peripheral tolerance are also compatible with means-directed negative selection, although means-directed central tolerance leaves little to do for peripheral tolerance. Moreover, the considerations leading up to Eq. (36) suggest that peripheral tolerance renders means-directed central tolerance all but superfluous (the efficacy of means-directed selection would rely entirely on the correlation between low clonotype means and low across-APP variances). In sharp contrast, central and peripheral tolerance would complement each other perfectly if the former were variance-directed. The latter maximises the efficacy of the repertoire by adjusting the sensitivity of each T cell to the level allowed by its across-APP variance, whereas central tolerance deletes most of the T cells which would be rendered effectively useless by peripheral tolerance (a similar idea is expressed in Grossman and Singer, 1996). Finally, pruning the repertoire of clones with high σ'^2_{APP}-values may enhance the reliability of an instructive tolerance mechanism.

CD4 T cells may take on the suppressor phenotype either in the thymus or in the periphery, and it is possible that this difference defines two functionally distinct subsets (Roncarolo and Levings, 2000; Seddon and Mason, 2000; Shevach, 2000; Jordan *et al.*, 2001). Variance-directed negative selection predicts such a difference. Thymus-educated suppressor cells are, by hypothesis, selected for high peripheral variances, and therefore autoreactive to "focal" autoantigens, restricted in their expression to a specific cell type, but occurring at high copy numbers where expressed. This implies that thymus-educated suppressor T cells would be particularly effective in averting organ-specific autoimmune diseases (cf. Egwuagu *et al.,* 1997; Itoh *et al.*, 1999). By contrast, it follows from variance-directed negative selection that peripherally-educated suppressor cells are more generally autoreactive, registering a relatively high TCR signal on average across peripheral APPs. Thus peripherally-educated suppressor T cells may play a more generic role in modulating immune responses (cf. Seddon and Mason, 1999; Roncarolo and Levings, 2000).

DISCUSSION

The need for peripheral tolerance mechanisms (deletion, anergy, regulation) has traditionally been explained from the inability of thymic cells to express autoantigens specific to other cell lineages, which renders central tolerance unable to purge the T cell repertoire completely of autoreactivity (Kappler *et al.*, 1987; Ardavín *et al.*, 1993; Sospreda *et al.*, 1998). This argument seemed particularly persuasive in the case of class I presentation, which, barring cross-presentation, relies on intracellular expression of the proteins from which the epitopes are derived. Class II presentation appeared less complex, because expression of MHC class II molecules is restricted to thymic epithelium and pAPCs, and presentation on these molecules usually follows internalisation of the proteins that yield the epitopes. Thus, both tolerance and potential autoimmunity should be determined primarily by the concentration of circulating protein (which can access the thymus; Kyewski *et al.*, 1984) rather than the cell type of origin (Klein *et al.*, 1998). None the less, even for class II a problem analogous to that of class I arises in the case of inducible proteins, which would not normally be found in the circulation during thymic selection (Klein *et al.*, 1998). Thus, for both classes the need for peripheral tolerance could be argued from an assumed thymic inability to present all potentially autoimmunogenic epitopes. However, this thymic inability is not absolute, as is attested by the evidence that many rare, inducible, or sequestered epitopes may in fact be presented during thymic selection (Farr and Rudensky, 1998; Hanahan, 1998; Klein *et al.*, 1998; Sospreda *et al.*, 1998). While these findings have seemingly made it more difficult to account for the co-existence of central and peripheral tolerance, the ability to determine thymic presentation patterns empirically also suggests a more satisfying resolution of the problem, since these patterns effectively encode the target of negative selection. The present paper shows that different hypotheses on the target of negative selection lead to different predictions concerning thymic presentation. Once this role has been established, its relative importance among peripheral tolerance mechanisms can be evaluated. As shown by the example of variance-directed negative selection, central and peripheral tolerance can perfectly complement each other.

The main message of this paper is that the target of negative selection can be determined objectively from presentation statistics on the negatively selecting cells and

the pAPCs in the secondary lymphoid tissues. Thus, the question of whether or not negative selection is variance-directed rather than means-directed (to name a specific example) are shown here to be amenable to experimental analysis, since certain differences in thymic presentation of autoantigens would count as evidence in favour of the former.

Further, indirect evidence for the variance-directed negative selection hypothesis would be the confirmation of the prediction that thymus-educated suppressor cells (which have a high σ'^2_{APP}-value) should respond to tissue-specific self, whereas peripherally-educated suppressor cells (which have a high $\bar{w}_i$ value) should respond to ubiquitous self. The former subset might correspond to the $CD25^+CD4^+$ regulatory T cells (Jordan *et al.*, 2001), the latter to the Tr1 $CD4^+$ cells (Roncarolo and Levings, 2000).

The expressions for P_{surv} presented in "The survival probability" do not take into account an additional role of negative selection, which is to delete thymocytes whose TCR is able to bind two or more MHC isoforms expressed by the thymic cells (Ignatowicz *et al.*, 1996; Zerrahn *et al.*, 1997; van den Berg and Rand, 2003). Deletion due to lack of MHC restriction, as opposed to deletion due to specific autoreactivity, may account for much of the observed death rate due to negative selection (Surh and Sprent, 1994; Ignatowicz *et al.*, 1996; Meerwijk *et al.*, 1997; Merkenschlager *et al.*, 1997; Bouneaud *et al.*, 2000). However, a minor quantitative contribution to the thymic death rate does not imply an unimportant contribution to tolerance: even a minute autoreactive fraction of the repertoire already poses a significant danger of autoimmunity. It is even possible that thymocyte death due to negative selection is entirely due to low-affinity recognition of more than one MHC isoform, and that specific autorecognition always induces differentiation to the suppressor phenotype rather than apoptosis (cf. Seddon and Mason (2000); Shevach (2000)). If this intriguing scenario is correct, the quantity P_{surv} as used in the present theory should be interpreted as the probability of maturing into a naïve main repertoire cell rather than into a suppressor T cell.

It is difficult to establish experimentally the number of encounters with a negatively-selecting APC that an individual thymocyte must undergo (Ardavín *et al.*, 1993; Merkenschlager *et al.*, 1994; Scollay and Godfrey, 1995). The present theory relates this number m to intrathymic presentation patterns. A maximal effect obtains for $m \gg m_{\mathrm{crit}}$, and the critical number of rounds can be calculated from the effective mean ubiquity, which can be determined in principle, by estimating the fraction of negatively-selecting cells presenting the autoantigen at any given moment (Smith *et al.*, 1997).

In establishing the efficiency principle we have considered statistics of the form $\langle \boldsymbol{\alpha}, \hat{\boldsymbol{L}} \rangle$ only. However, the method can be extended to any real scalar statistic of the form $t(\boldsymbol{\nu})$; the general problem is to seek a $\boldsymbol{\rho}_\omega$ such that $\mathbb{P}\{t(\boldsymbol{\nu}) > \omega \,|\, m \,\text{rounds}\}$ is minimised. Also, the restriction to a Bernoulli distribution for w_{ij} is not essential, and generalisation of the present results need not pose major difficulties. A distribution which has a mass between the extremes of no recognition and maximum recognition accounts for so-called "avidity" differences between clonotypes responding to a given antigen. A critical question in this regard is whether the activation curves as shown in Fig. 4 retain their plateau at μ; numerical calculations based on a continuous distribution for the TCR signal $w_{i\boldsymbol{\zeta}}$ suggest that this crucial feature is preserved (van den Berg *et al.*, 2001).

Two additional sources of stochasticity in TCR signalling were ignored in the present analysis. The first is MHC loading fluctuations: strictly speaking, ζ_j is the *probability* that a pMHC complex will be of species j, and the realised APP is a multinomial variate. However, these complications can be safely ignored if the MHC density is high (appendix C in van den Berg *et al.*, 2001). The second source of variation is inherent in the TCR signalling process, which is based on TCR ligation events, which follow chance encounters of TCR and pMHC molecules. This variability can be ignored if the conjugate is sufficiently long-lasting relative to the typical ligation time of a good agonist (incidentally, this provides an explanation for the fact that the latter time scale is relatively short). These conditions of high MHC density and long conjugation time are typically met by the professional antigen-presenting cells involved in the selection of thymocytes and the activation of naïve T cells (Iezzi *et al.*, 1998). However, the situation may be dramatically different for CTLs operating in the tissues, where conjugation times may be short and MHC densities on infected cells may be low (Janeway and Travers, 1997).

Tolerance has been discussed in terms of clonal selection ever since McFarlane Burnet first coined the phrase (Burnet, 1959). However, as will be clear from the theoretical development in this paper, the natural unit of negative selection is the individual thymocyte, not the clone. (The importance of making this distinction depends on the number of thymocytes that belong to the same clone, since the probability that the clone will be entirely eliminated diminishes exponentially with this number.) Similarly, the individual mature T cell is the natural object of peripheral tolerance and of activation during an immune response. Both are governed by the T cell's law, which remains a clonal characteristic inherited from the progenitor of the clone to which the T cell belongs.

Central tolerance can shape the TCR repertoire, specifically geared towards complementing the role of peripheral tolerance, by having distinct thymic autoantigen presentation statistics that are altered in a particular way relative to those of the secondary lymphoid tissues. Of course, it is difficult to envisage a mechanism whereby the selecting cells in the thymus somehow "know" the peripheral presentation statistics, and "decide" to adapt their own presentation in a specific way. But this is not needed. All that is required are slight changes in thymic presentation (here represented by $\boldsymbol{\rho}$)

by mutations causing ectopic thymic expression of certain autoantigens. Provided that the contribution made by central tolerance improves smoothly as $\boldsymbol{\rho}$ converges toward the "true" target $\tilde{\boldsymbol{\alpha}}$, natural selection can act to effect this convergence. The "true" target $\tilde{\boldsymbol{\alpha}}$ need not be known explicitly by the system, but is determined by the particular properties of peripheral presentation and tolerance mechanisms.

Acknowledgements

The authors gratefully acknowledge the insights and comments contributed by Don Mason, Andrew Sewell, Nigel Burroughs, David Rand, and two referees.

References

Akkaraju, S., Ho, W.Y., Leong, D., Canaan, K., Davis, M.M. and Goodnow, C.C. (1997) "A range of CD4 T cell tolerance: partial inactivation to organ-specific antigen allows nondestructive thyroiditis or insulitis", *Immunity* **7**, 255–271.

Anderson, A.C., Waldner, H., Turchin, V., Jabs, C., Prabhu Das, M., Kuchroo, V.K. and Nicholson, L.B. (2000) "Autoantigen-responsive T cell clones demonstrate unfocused TCR cross-reactivity toward multiple related ligands: implications for autoimmunity", *Cell. Immunol.* **202**, 88–96.

Antonia, S.J., Geiger, T., Miller, J. and Flavell, R.A. (1995) "Mechanisms of immune tolerance induction through the thymic expression of a peripheral tissue-specific protein", *Int. Immunol.* **7**, 715–725.

Ardavín, C., Wu, L., Li, C.-L. and Shortman, K. (1993) "Thymic dendritic cells and T cells develop simultaneously in the thymus from a common precursor population", *Nature* **362**, 761–763.

Arstila, T.P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J. and Kourilsky, P. (1999) "A direct estimate of the human $\alpha - \beta$ T cell receptor diversity", *Science* **286**, 958–961.

van den Berg, H.A. and Rand, D.A. (2003) "Antigen presentation on MHC molecules as a diversity filter that enhances immune efficacy", *J. Theor. Biol.* **224**, 249–267.

van den Berg, H.A., Rand, D.A. and Burroughs, N.J. (2001) "A reliable and safe T cell repertoire based on low-affinity T cell receptors", *J. Theor. Biol.* **209**, 465–486.

van den Berg, H.A., Burroughs, N.J. and Rand, D.A. (2002) "Quantifying the strength of ligand antagonism in TCR triggering", *Bull. Math. Biol.* **64**, 781–808.

Bevan, M.J., Hogquist, K.A. and Jameson, S.C. (1994) "Selecting the T cell repertoire", *Science* **264**, 796–797.

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

Bouneaud, C., Kourilsky, P. and Bousso, P. (2000) "Impact of negative selection on the T cell repertoire reactive to a self-peptide: a large fraction of T cell clones escapes clonal deletion", *Immunity* **13**, 829–840.

Burnet, F.M. (1959) The Clonal Selection Theory of Acquired Immunology (Cambridge University Press, Cambridge).

Butz, E.A. and Bevan, M.J. (1998) "Massive expansion of antigen-specific $CD8^+$ T cells during an acute virus infection", *Immunity* **8**, 167–175.

Byers, D.E. and Lindahl, K.F. (1999) "Peptide affinity and concentration affect the sensitivity of M3-restricted CTLs *in vitro*", *J. Immunol.* **163**, 3022–3028.

Davis, M.M., Boniface, J.J., Reich, Z., Lyons, D., Hampl, J., Arden, B. and Chien, Y.-h. (1998) "Ligand recognition by $\alpha\beta$ T cell receptors", *Annu. Rev. Immunol.* **16**, 523–544.

Dembo, A. and Zeitouni, O. (1998) Large Deviations Techniques and Applications (Springer Verlag, New York).

Egwuagu, C.E., Charukamnoetkanok, P. and Gery, I. (1997) "Thymic expression of autoantigens correlates with resistance to autoimmune disease", *J. Immunol.* **159**, 3109–3112.

Farr, A.G. and Rudensky, A. (1998) "Medullary thymic epithelium: a mosaic of epithelial self?", *J. Exp. Med.* **188**, 1–4.

Fritz, R.B. and Zhao, M.-L. (1996) "Thymic expression of myelin basic protein (MBP)—activation of MBP-specific T cells by thymic cells in the absence of exogenous MBP", *J. Immunol.* **157**, 5249–5253.

Gallucci, S. and Matzinger, P. (2001) "Danger signals: SOS to the immune system", *Curr. Opin. Immunol.* **13**, 114–119.

Gavin, M.A. and Bevan, M.J. (1995) "Increased peptide promiscuity provides a rationale for the lack of N regions in the neonatal T cell repertoire", *Immunity* **3**, 793–800.

Grossman, Z. and Paul, W.E. (1992) "Adaptive cellular interactions in the immune system: the tunable activation threshold and the significance of subthreshold responses", *Proc. Natl Acad. Sci. USA* **89**, 10365–10369.

Grossman, Z. and Paul, W.E. (2001) "Autoreactivity, dynamic tuning and selectivity", *Curr. Opin. Immunol.* **13**, 687–698.

Grossman, Z. and Singer, A. (1996) "Tuning of activation thresholds explains flexibility in the selection and development of T cells in the thymus", *Proc. Natl Acad. Sci. USA* **93**, 14747–14752.

Hagerty, D.T. and Allen, P.M. (1995) "Intramolecular mimicry: identification and analysis of two cross-reactive T cell epitopes within a single protein", *J. Immunol.* **155**, 2993–3001.

Hanahan, D. (1998) "Peripheral-antigen-expressing cells in thymic medulla: factors in self-tolerance and autoimmunity", *Curr. Opin. Immunol.* **10**, 214–219.

Hawiger, D., Inaba, K., Dorsett, Y., Guo, M., Mahnke, K., Rivera, M., Ravetch, J.V., Steinman, R.M. and Nussenzweig, M.C. (2001) "Dendritic cells induce peripheral T cell unresponsiveness under steady state conditions *in vivo*", *J. Exp. Med.* **194**, 769–779.

Hernández, J., Lee, P.P., Davis, M.M. and Sherman, L.A. (2000) "The use of HLA A2.1/p53 peptide tetramers to visualize the impact of self tolerance on the TCR repertoire", *J. Immunol.* **164**, 596–602.

Hogquist, K.A., Tomlinson, A.J., Kieper, W.C., McGargill, M.A. and Hart, M.C. (1997) "Identification of a naturally occurring ligand for thymic positive selection", *Immunity* **6**, 389–399.

Hudrisier, D., Kessler, B., Valitutti, S., Horvath, C., Cerottini, J.-C. and Luescher, I.F. (1998) "The efficiency of antigen recognition by $CD8^+$ CTL clones is determined by the frequency of serial TCR engagement", *J. Immunol.* **161**, 553–562.

Hunt, D.F., Henderson, R.A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A.L., Appella, E. and Engelhard, V.H. (1992) "Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry", *Science* **255**, 1261–1263.

Iezzi, G., Karjalainen, K. and Lanzavecchia, A. (1998) "The duration of antigenic stimulation determines the fate of naive and effector T cells", *Immunity* **8**, 89–95.

Ignatowicz, L., Kappler, J. and Marrack, P. (1996) "The repertoire of T cells shaped by a single MHC/peptide ligand", *Cell* **84**, 521–529.

Ignatowicz, L., Rees, W., Pacholczyk, R., Ignatowicz, H., Kushnir, E., Kappler, J. and Marrack, P. (1997) "T cells can be activated by peptides that are unrelated in sequence to their selecting peptide", *Immunity* **7**, 179–186.

Itoh, Y. and Germain, R.N. (1997) "Single cell analysis reveals regulated hierarchical T cell antigen receptor signaling thresholds and intraclonal heterogeneity for individual cytokine responses of $CD4^+$ cells", *J. Exp. Med.* **186**, 757–766.

Itoh, M., Takahashi, T., Sakaguchi, N., Kuniyasu, Y., Shimizu, J., Otsuka, F. and Sakaguchi, S. (1999) "Thymus and autoimmunity: production of $CD35^+CD^+$ naturally anergic and suppressive T cells as a key function of the thymus in maintaining immunologic self-tolerance", *J. Immunol.* **162**, 5317–5326.

Janeway, C.A. Jr. and Travers, P. (1997) Immunobiology: The Immune System in Health and Disease, 3rd Ed. (Garland Publishing, New York).

Jardetzky, T.S., Lane, W.S., Robinson, R.A. and Wiley, D.C. (1991) "Identification of self peptides bound to purified HLA-B27", *Nature* **353**, 326–329.

Jordan, M.S., Boesteanu, A., Reed, A.J., Petrone, A.L., Holenbeck, A.E., Lerman, M.A., Naji, A. and Caton, A.J. (2001) "Thymic selection of $CD4^+CD25^+$ regulatory T cells induced by an agonist self-peptide", *Nature Immunol.* **2**, 301–306.

Kalergis, A.M., Boucheron, N., Doucey, M.-A., Palmieri, E., Goyarts, E.C., Vegh, Z., Luesher, I.F. and Nathenson, S.G. (2001) "Efficient T cell activation requires an optimal dwell-time of interaction between the TCR and the pMHC complex", *Nature Immunol.* **2**, 229–234.

Kappler, J.W., Roehm, N. and Marrack, P. (1987) "T cell tolerance by clonal elimination in the thymus", *Cell* **49**, 273–280.

Kisielow, P., Blüthmann, H., Staerz, U.D., Steinmetz, M. and von Boehmer, H. (1988) "Tolerance in T-cell-receptor transgenic

mice involves deletion of nonmature $CD4^+8^+$ thymocytes", *Nature* **333**, 742–746.
Klein, L., Klein, T., Rüther, U. and Kyewski, B. (1998) "CD4 T cell tolerance to human C-reactive protein, an inducible serum protein, is mediated by medullary thymic epithelium", *J. Exp. Med.* **188**, 5–16.
Kurts, C., Sutherland, R.M., Davey, G., Lew, A.M., Blanas, E., Carbone, F.R., Miller, J.F.A.P. and Heath, W.R. (1999) "CD8 T cell ignorance or tolerance to islet antigens depends on antigen dose", *Proc. Natl Acad. Sci. USA* **96**, 12703–12707.
Kyewski, B.A., Fathman, C.G. and Kaplan, H.S. (1984) "Intrathymic presentation of circulating non-major histocompatibility antigens", *Nature* **308**, 196–199.
Lanzavecchia, A. and Sallusto, F. (2000) "Dynamics of T lymphocyte responses: intermediates, effectors, and memory cells", *Science* **290**, 92–97.
Laufer, T.M., Glimcher, L.H. and Lo, D. (1999) "Using thymus anatomy to dissect T cell repertoire selection", *Semin. Immunol.* **11**, 65–70.
Lombardi, G., Sidhu, S., Batchelor, R. and Lechler, R. (1994) "Anergic T cells as suppressor cells *in vitro*", *Science* **264**, 1587–1589.
Mason, D. (1998) "A very high level of crossreactivity is an essential feature of the T-cell receptor", *Immunol. Today* **19**, 395–404.
Mason, D. (2001) "Some quantitative aspects of T cell repertoire selection: the requirement for regulatory T cells", *Immunol. Rev.* **182**, 80–88.
Mazza, G., Housset, D., Piras, C., Gregoire, C., Lin, S.-Y., Fontecilla-Camps, J.C. and Malissen, B. (1998) "Glimpses at the recognition of peptide/MHC complexes by T-cell antigen receptors", *Immunol. Rev.* **163**, 187–196.
Meerwijk, J.P.M., van Marguerat, S., Lees, R.K., Germain, R.N., Fowlkes, B.J. and MacDonald, H.R. (1997) "Quantitative impact of thymic clonal deletion on the T cell repertoire", *J. Exp. Med.* **185**, 377–383.
Merkenschlager, M. (1996) "Tracing interactions of thymocytes with individual stromal cell partners", *Eur. J. Immunol.* **26**, 892–896.
Merkenschlager, M., Benoist, C. and Mathis, D. (1994) "Evidence for a single-niche model of positive selection", *Proc. Natl Acad. Sci. USA* **91**, 11694–11698.
Merkenschlager, M., Graf, D., Lovatt, M., Bommhardt, U., Zamoyska, R. and Fisher, A.G. (1997) "How many thymocytes audition for selection?", *J. Exp. Med.* **186**, 1149–1158.
Morgan, D.J., Kreuwel, H.T.C. and Sherman, L.A. (1999) "Antigen concentration and precursor frequency determine the rate of $CD8^+$ T cell tolerance to peripherally expressed antigens", *J. Immunol.* **163**, 723–727.
Nicholson, L.B., Anderson, A.C. and Kuchroo, V.K. (2000) "Tuning T cell activation threshold and effector function with cross-reactive peptide ligands", *Int. Immunol.* **12**, 205–213.
Nugent, C.T., Morgan, D.J., Biggs, J.A., Ko, A., Pilip, I.M., Pamer, E.G. and Sherman, L.A. (2000) "Characterization of $CD8^+$ T lymphocytes that persist after peripheral tolerance to a self antigen expressed in the pancreas", *J. Immunol.* **164**, 191–200.
Oelke, M., Maus, M.V., Didiano, D., June, C.H., Mackensen, A. and Schneck, J.P. (2003) "*Ex vivo* induction and expansion of antigen-specific cytotoxic T cells by HLA-Ig coated artificial antigen-presenting cells", *Nature Med.* **9**, 619–625.
Pittet, M.J., Valmori, D., Dunbar, P.R., Speiser, D.E., Liénard, D., Lejeune, F., Fleischhauer, K., Cerundolo, V., Cerottini, J.-C. and Romero, P. (1999) "High frequencies of naive Melan-A/MART-1-specific $CD8^+$ T cells in a large proportion of human Histocompatibility Leukocyte Antigen (HLA)-A2 individuals", *J. Exp. Med.* **190**, 705–715.
Reay, P.A., Matsui, K., Haase, K., Wülfing, C., Chien, Y.-H. and Davis, M. (2000) "Determination of the relationship between T cell responsiveness and the number of MHC-peptide complexes using specific monoclonal antibodies", *J. Immunol.* **164**, 5626–5634.
Roncarolo, M.-G. and Levings, M.K. (2000) "The role of different subsets of T regulatory cells in controlling autoimmunity", *Curr. Opin. Immunol.* **12**, 676–683.
Savage, P.A. and Davis, M.M. (2001) "A kinetic window constricts the T cell receptor repertoire in the thymus", *Immunity* **14**, 243–252.
Scollay, R. and Godfrey, D.I. (1995) "Thymic emigration: conveyor belts or lucky dips?", *Immunol. Today* **16**, 268–273.
Sebza, E., Mariathasan, S., Ohteki, T., Jones, R., Bachmann, M.F. and Ohashi, P.S. (1999) "Selection of the T cell repertoire", *Annu. Rev. Immunol.* **17**, 829–874.
Seddon, B. and Mason, D. (1999) "Peripheral autoantigen induces regulatory T cells that prevent autoimmunity", *J. Exp. Med.* **189**, 877–881.
Seddon, B. and Mason, D. (2000) "The third function of the thymus", *Immunol. Today* **21**, 95–99.
Shevach, E.M. (2000) "Regulatory T cells in autoimmunity", *Annu. Rev. Immunol.* **18**, 423–449.
Smith, K.M., Olson, D.C., Hirose, R. and Hanahan, D. (1997) "Pancreatic gene expression in rare cells of thymic medulla: evidence for functional contribution to T cell tolerance", *Int. Immunol.* **9**, 1355–1365.
Sospreda, M., Ferrer-Francesch, X., Domúnguez, O., Juan, M., Foz-Sala, M. and Pujol-Borrell, R. (1998) "Transcription of a broad range of self-antigens in human thymus suggests a role for central mechanisms in tolerance toward peripheral antigens", *J. Immunol.* **161**, 5918–5929.
Steinman, R.M., Turley, S. and Mellman, I. (2000) "The induction of tolerance by dendritic cells that have captured apoptotic cells", *J. Exp. Med.* **191**, 411–416.
Stevanović, S. and Schild, H. (1999) "Quantitative aspects of T cell activation—peptide generation and editing by MHC class I molecule", *Semin. Immunol.* **11**, 375–384.
Surh, C.D. and Sprent, J. (1994) "T-cell apoptosis detected *in situ* during positive and negative selection in the thymus", *Nature* **372**, 100–103.
Tanchot, C., Lemonnier, F.A., Pérarnau, B., Freitas, A.A. and Rocha, B. (1997) "Differential requirements for survival and proliferation of CD8 naïve or memory T cells", *Science* **276**, 2057–2062.
Valitutti, S., Müller, S., Dessing, M. and Lanzavecchia, A. (1996) "Different responses are elicited in cytotoxic T lymphocytes by different levels of T cell receptor occupancy", *J. Exp. Med.* **183**, 1917–1921.
Viola, A. and Lanzavecchia, A. (1996) "T cell activation determined by T cell receptor number and tunable thresholds", *Science* **273**, 104–106.
de Visser, K.E., Cordaro, T.A., Kioussis, D., Haanen, J.B.A.G., Schumacher, T.N.M. and Kruisbeek, A.M. (2000) "Tracing and characterization of the low-avidity self-specific T cell repertoire", *Eur. J. Immunol.* **30**, 1458–1468.
Webb, S., Morris, C. and Sprent, J. (1990) "Extrathymic tolerance of mature T cells: clonal elimination as a consequence of immunity", *Cell* **63**, 1249–1256.
Wong, P., Barton, G.M., Forbush, K.A. and Rudensky, A.Y. (2001) "Dynamic tuning of T cell reactivity by self-peptide-major histocompatibility complex ligands", *J. Exp. Med.* **193**, 1179–1187.
Zerrahn, J., Held, W. and Raulet, D.H. (1997) "The MHC reactivity of the T cell repertoire prior to positive and negative selection", *Cell* **88**, 627–636.
Zippelius, A., Pittet, M., Batard, P., Rufer, N., de Smedt, M., Guillaume, P., Ellefsen, K., Valmori, D., Liénard, D., Plum, J., MacDonald, H.R., Speiser, D.E., Cerrottini, J.-C. and Romero, P. (2002) "Thymic selection generates a large T cell pool recognizing a self-peptide in humans", *J. Exp. Med.* **195**, 485–494.

A STATISTICS OF THE TCR SIGNAL

It is not immediately obvious that the Bernoulli approximation of Eq. (2) is reasonable, since mean dwell-times may be expected on physical grounds to be continuously distributed. This section offers a heuristic justification, based on a model of TCR triggering kinetics which has been formulated and analysed elsewhere (van den Berg *et al.*, 2002).

The hypothesis of the TCR triggering model is that a TCR-pMHC interaction evokes a signalling event whenever the TCR/pMHC ternary complex stays together for longer than threshold duration T_R (which is of the order of several seconds; Davis *et al.*, 1998; Hudrisier *et al.*, 1998). It can then be shown that the time-average rate of TCR triggering is $M_T \sum_j \zeta_j \tilde{w}_{ij}$, with MHC-specific

triggering rate

$$\tilde{w}_{ij} = \frac{[R]}{[R] + K_{D,ij}} \exp\{-T_R/T_{ij}\}/T_{ij}. \tag{A.1}$$

Here $[R]$ is the surface density of free (unoccupied) TCR molecules, $K_{D,ij}$ is the two-dimensional dissociation constant, characteristic of the complex consisting of TCR i and pMHC j, and T_{ij} is the mean dwell-time of the ternary complex. Thus, when $[R] \gg K_{D,ij}$, the mean dwell-time T_{ij} is the sole determinant of the MHC-specific triggering rate $\tilde{w}_{ij}$, with optimal ligands having a T_{ij} (nearly) equal to T_R (cf. Kalergis *et al.* (2001)). Otherwise, while T_{ij} is still important, $\tilde{w}_{ij}$ also depends on the affinity of the interaction, as well as on $[R]$ which itself depends on the APP and the affinities of all presented pMHC species. Thus the most potent peptide ligands combine high affinity (low $K_{D,ij}$) with a T_{ij} near T_R (see van den Berg *et al.*, 2002 for a detailed account). The following analysis is restricted to the case where $[R] \gg K_{D,ij}$. The MHC-specific triggering rate is then given by:

$$\tilde{w}_{ij} = \frac{\exp\{-T_R/T_{ij}\}}{T_{ij}} \tag{A.2}$$

so that $\tilde{w}_{ij}$ attains its maximum at $(eT_R)^{-1}$ for $T_{ij} = T_R$. Thus, $w_{ij} \stackrel{\text{def}}{=} \tilde{w}_{ij} eT_R$ is the normalized TCR triggering rate introduced in Eqs. (1) and (2).

To motivate the Bernoulli approximation, let $f_T(t)$ denote the mass function of the mean dwell-time T_{ij}, taken to be continuous with $f_T(t) > 0$ for $t \in (0, T_R + \delta)$ for some $\delta > 0$, without an essential singularity at $t = 0$. The distribution function of the TCR triggering rate is

$$F_W(\omega) \stackrel{\text{def}}{=} \mathbb{P}(\tilde{w}_{ij} \leq \omega) = 1 - \int_{T_1(\omega)}^{T_2(\omega)} f_T(t)\mathrm{d}t$$

where T_1 and T_2 are found by solving Eq. (A.2) for T_{ij}, with $\tilde{w}_{ij} = \omega$. The associated mass function is

$$F'_W(\omega) = \frac{T_1^3 f_T(T_1) \exp\{T_R/T_1\}}{T_R - T_1} - \frac{T_2^3 f_T(T_2) \exp\{T_R/T_2\}}{T_R - T_2}$$

by Leibnitz's rule. It is clear from Eq. (A.2) that T_1 and T_2 approach T_R, from below and above, respectively, as ω approaches the maximum value $(eT_R)^{-1}$ from below. Then

$$\begin{aligned}\lim_{\omega \to [(eT_R)^{-1}]^-} & F'_W(\omega) \\ &= T_R^3 e^1 f_T(T_R) \left(\lim_{T_1 \to [T_R]^-} \frac{1}{T_R - T_1} + \lim_{T_2 \to [T_R]^+} \frac{1}{T_2 - T_R} \right) \\ &= +\infty. \end{aligned} \tag{A.3}$$

Next, observe from Eq. (A.2) that

$$T_1(\omega) \to 0^+ \quad \text{and} \quad T_2(\omega) \to +\infty \quad \text{as} \quad \omega \to 0^+.$$

By assumption, $f_T(t)$ has no essential singularity at $t = 0$, and therefore

$$\begin{aligned}\lim_{T_1 \to 0^+} & \frac{T_1^3 f_T(T_1) \exp\{T_R/T_1\}}{T_R - T_1} \\ &= \lim_{T_1 \to 0^+} \frac{\exp\{3\ln\{T_1\} + \ln\{f_T(T_1)\} + T_R/T_1\}}{T_R - T_1} \\ &= +\infty; \end{aligned}$$

and since

$$\begin{aligned}\lim_{T_2 \to \infty} & \frac{T_2^3 f_T(T_2) \exp\{T_R/T_2\}}{T_R - T_2} \\ &= -\lim_{T_2 \to \infty} \exp\{3\ln\{T_2\} - \ln\{T_2 - T_R\} \\ &\quad + \ln\{f_T(T_2)\} + T_R/T_2\} \end{aligned}$$

is either finite or tends to $-\infty$, it follows in either case that

$$\lim_{\omega \to 0^+} F'_W(\omega) = +\infty. \tag{A.4}$$

Together, Eqs. (A.3) and (A.4) show that the TCR triggering rate formula, Eq. (A.2), concentrates probability at the minimum and maximum triggering rate, even if the T_{ij} follow a continuous distribution. These considerations provide an heuristic motivation for the Bernoulli approximation of Eq. (2).

B LARGE DEVIATIONS RATE FUNCTIONS

The pre-selection repertoire statistics is determined by the entropy of each clonotype's law $\hat{\boldsymbol{L}}_i$ relative to the probability μ of recognising any given pMHC species. This entropy $H(\hat{\boldsymbol{L}}|\mu)$ is derived in "Across-clonotype statistics". Clonotype entropies are modified by negative selection; calculation of this modified entropy

$$H(\hat{\boldsymbol{L}}|\mu) + m\left\{ \inf_{\boldsymbol{w}:\langle \mathbf{1}, \boldsymbol{w} \rangle \leq w_{\text{thy}}} \langle \boldsymbol{\pi}, \boldsymbol{I}(\boldsymbol{w}; \hat{\boldsymbol{L}}, \boldsymbol{u}, \boldsymbol{\rho}) \rangle \right\} - I_{\text{surv}}$$

(the infimand in Eq. (14)) requires the large deviations rate function for across-APP variations, given a clonotype law $\hat{\boldsymbol{L}}$, which is derived in "Within-clonotype, across-APP, statistics". The calculation of the post-selection large deviations rate function from the modified entropy is discussed in "The post-selection large deviations rate function".

B.1 Across-clonotype Statistics

Fix a clonotype i, and consider the probability $p_N(i,k)$ that a pMHC species j, chosen at random, belongs to component k and is recognised by a TCR of clonotype i:

$$p_N(i,k) \stackrel{\text{def}}{=} \mathbb{P}\{j \in k \,\&\, w_{ij} = 1\}; \tag{B.1}$$

the probability that the pMHC belongs to component k and is not recognised is $\pi_k - p_N(i,k)$. The probability prior to negative selection that the clonotype mean TCR signal across antigen presentation patterns $\bar{w}_i$ equals ω satisfies, by Sanov's theorem (Dembo and Zeitouni, 1998), the following large deviations principle:

$$\lim_{N\to\infty} \frac{1}{N} \ln \mathbb{P}(\bar{w}_i = \omega) = -\inf_{\boldsymbol{p}\in\{\boldsymbol{p}:\bar{w}_i=\omega\}} H(\boldsymbol{p}|\mu) \tag{B.2}$$

where $\boldsymbol{p} \in [0,1]^K$ and $H(\boldsymbol{p}|\mu)$ expresses the *relative entropy* of $\boldsymbol{p}$ relative to μ:

$$H(\boldsymbol{p}|\mu) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \left[p_k \ln \frac{p_k}{\pi_k \mu} + (\pi_k - p_k) \ln \frac{\pi_k - p_k}{\pi_k(1-\mu)} \right]. \tag{B.3}$$

Now $\hat{L}_{ik} = p_N(i,k)/\pi_k$ denotes the recognition frequency of clonotype i for component i. In terms of these recognition frequencies, the relative entropy becomes, for a given clonotype law $\hat{\boldsymbol{L}}_i \in \mathscr{L}_N$:

$$H(\hat{\boldsymbol{L}}_i|\mu) = \sum_{k=1}^{K} \pi_k \left(\hat{L}_{ik} \ln \frac{\hat{L}_{ik}}{\mu} + (1 - \hat{L}_{ik}) \ln \frac{1-\hat{L}_{ik}}{1-\mu} \right) \tag{B.4}$$

and Sanov's result becomes

$$\lim_{N\to\infty} \frac{1}{N} \ln \mathbb{P}(\bar{w}_i = \omega) = -\inf_{\boldsymbol{\nu}\in\{\boldsymbol{\nu}:\langle\boldsymbol{\rho},\boldsymbol{\nu}\rangle=\omega\}} H(\boldsymbol{\nu}|\mu). \tag{B.5}$$

B.2 Within-clonotype, across-APP, Statistics

This section aims to establish a large deviations principle for the probabilities

$$\lim_{N\to\infty} \frac{1}{N} \ln \mathbb{P}\{w_{i\boldsymbol{\zeta}} \leqq \omega\} \quad \text{for} \quad \omega \leqq \bar{w}_i,$$

given that the T cell at hand belongs to clonotype i with recognition frequencies $\hat{\boldsymbol{L}}_i$, and clonotype mean $\bar{w}_i = \langle \boldsymbol{\rho}, \hat{\boldsymbol{L}}_i \rangle$. Fluctuations across the APPs for a given clonotype with law $\hat{\boldsymbol{L}}_i$ are represented by the random vector $\boldsymbol{Y}_{i\boldsymbol{\zeta}}$, whose elements are defined by

$$Y_{ki\boldsymbol{\zeta}} \stackrel{\text{def}}{=} \frac{\hat{n}_{ik\boldsymbol{\zeta}}}{N_k}$$

where $\hat{n}_{ik\boldsymbol{\zeta}}$ denotes the number of component k pMHC species present in APP $\boldsymbol{\zeta}$ that are recognised by a TCR of clonotype i. This number is hypergeometrically distributed, Eq. (10). The number of pMHC species from component k and present in APP $\boldsymbol{\zeta}$, defined as $n_{k\boldsymbol{\zeta}} \stackrel{\text{def}}{=} \sum_{j\in\mathscr{K}_k} I_{k\boldsymbol{\zeta}}$, may for large N be replaced by its expectation $u_k N_k$, and thus

$$\mathbb{P}(Y_{ik\boldsymbol{\zeta}} = y) = \frac{\begin{pmatrix} \hat{L}_{ik} N_k \\ y N_k \end{pmatrix} \begin{pmatrix} (1-\hat{L}_{ik}) N_k \\ (u_k - y) N_k \end{pmatrix}}{\begin{pmatrix} N_k \\ u_k N_k \end{pmatrix}}. \tag{B.6}$$

The binomial coefficients in this equation can be estimated by Stirling's formula

$$\lim_{N\to\infty} \left\{ \frac{\ln\{N!\}}{N} - \ln\{N\} \right\} = -1$$

which leads to the estimate

$$\mathbb{P}(Y_{ik\boldsymbol{\zeta}} = y) \sim \exp\{ -N_k J(y; \hat{L}_{ik}, u_k)\}$$

where

$$\begin{aligned} J(y;\nu,u) \stackrel{\text{def}}{=}\ & y \ln\{y\} + (y - [\nu + u - 1]) \\ & \times \ln\{y - [\nu + u - 1]\} + (\nu - y) \ln\{\nu - y\} \\ & + (u - y) \ln\{u - y\} - (\nu \ln\{\nu\} + (1-\nu) \ln\{1-\nu\} \\ & + u \ln\{u\} + (1-u) \ln\{1-u\}). \end{aligned} \tag{B.7}$$

The function J is a steep rate function (Dembo and Zeitouni, 1998):

$$J(y;\nu,u) < +\infty \quad \text{for}$$

$$\max(0, \nu + u - 1) < y < \min(\nu, u),$$

and as y approaches these bounds, the derivative $\mathrm{d}J/\mathrm{d}y$ tends to $+\infty$. The bounds on y_k determine the minimum and maximum values which the TCR signal can assume:

$$\lfloor w_{i\boldsymbol{\zeta}} \rfloor \stackrel{\text{def}}{=} \sum_{k=1}^{K} \max(0, [\hat{L}_{ik} + u_k - 1]/u_k) \tag{B.8}$$

$$\lceil w_{i\boldsymbol{\zeta}} \rceil \stackrel{\text{def}}{=} \sum_{k=1}^{K} \min(\hat{L}_{ik}/u_k, 1). \tag{B.9}$$

Since

$$\Delta\mathbb{P}(Y_{ik\boldsymbol{\zeta}} \leq y) \approx \mathbb{P}(Y_{ik\boldsymbol{\zeta}} = y) N_k \Delta y,$$

the probability density function of $Y_{ik\boldsymbol{\zeta}}$ tends to $N_k \exp\{-N_k J(y; \hat{L}_{ik}, u_k)\}$ for large N_k. Thus for an

interval $\mathscr{I}$,

$$\lim_{N\to\infty}\frac{1}{N}\ln\mathbb{P}\{Y_{ik\boldsymbol{\zeta}}\in\mathscr{I}\}$$

$$=\pi_k\lim_{N_k\to\infty}\left[\frac{1}{N_k}\ln\int_{\mathscr{I}}\exp\{-N_kJ(y;\hat{L}_{ik},u_k)\}\mathrm{d}y\right]$$

$$=-\pi_k\inf_{y\in\mathscr{I}}J(y;\hat{L}_{ik},u_k) \tag{B.10}$$

by Laplace's principle of steepest descent. Let $\boldsymbol{I}(\boldsymbol{w};\hat{\boldsymbol{L}}_i,\boldsymbol{u},\boldsymbol{\rho})$ denote the function defined by

$$I_k(\boldsymbol{w};\hat{\boldsymbol{L}}_i,\boldsymbol{u},\boldsymbol{\rho})\stackrel{\text{def}}{=}J(w_ku_k/\rho_k;\hat{L}_{ik},u_k)\quad k=1,\ldots,K.$$

Here $\boldsymbol{w}$ is a K-vector collecting the contributions from the various components, so that the TCR signal is $\langle\mathbf{1},\boldsymbol{w}\rangle$. Repeated application of Varadhan's integral lemma (Dembo and Zeitouni, 1998) yields the desired asymptotic estimates:

$$\lim_{N\to\infty}\frac{1}{\mathrm{N}}\,\mathbb{P}\ln\{w_{i\boldsymbol{\zeta}}<\omega\}$$

$$=\begin{cases}0 & \omega\geq\bar{w}_i\\ -\inf_{\boldsymbol{w}\in\{\boldsymbol{w}:\langle\mathbf{1},\boldsymbol{w}\rangle=\omega\}}\langle\boldsymbol{\pi},\boldsymbol{I}(\boldsymbol{w};\hat{\boldsymbol{L}}_i,\boldsymbol{u},\boldsymbol{\rho})\rangle & \lfloor w_{i\boldsymbol{\zeta}}\rfloor<\omega<\bar{w}_i\\ -\infty & \omega<\lfloor w_{i\boldsymbol{\zeta}}\rfloor\end{cases} \tag{B.11}$$

which is used to calculate survival probabilities (with $\omega=w_{\text{thy}}$), and

$$\lim_{N\to\infty}\frac{1}{\mathrm{N}}\ln\mathbb{P}\{w_{i\boldsymbol{\zeta}}>\omega\}$$

$$=\begin{cases}0 & \omega<\bar{w}_i\\ -\inf_{\boldsymbol{w}\in\{\boldsymbol{w}:\langle\mathbf{1},\boldsymbol{w}\rangle=\omega\}}\langle\boldsymbol{\pi},\boldsymbol{I}(\boldsymbol{w};\hat{\boldsymbol{L}}_i,\boldsymbol{u},\boldsymbol{\rho})\rangle & \bar{w}_i<\omega<\lceil w_{i\boldsymbol{\zeta}}\rceil\\ -\infty & \omega\geq\lceil w_{i\boldsymbol{\zeta}}\rceil\end{cases} \tag{B.12}$$

which can be used to determine the autoimmunity probability (with $\omega=w_{\text{act}}$). The case where $\mu\ll1$ is immunologically relevant. Therefore, it may be assumed that $\hat{L}_{ik}\ll1$ for all components k, whence $1-\nu\doteq1$ and $u-y\doteq u$. The function J can be approximated as follows:

$$J(y,\nu,u)\doteq\nu\left(\frac{y}{\nu}\ln\frac{y/\nu}{u}+\left(1-\frac{y}{\nu}\right)\ln\frac{1-y/\nu}{1-u}\right). \tag{B.13}$$

The results in this section apply to both the pre- and post-selection situations, since the statistics take the law $\hat{\boldsymbol{L}}$ as parameter; and negative selection affects the distribution of the laws over the clonal repertoire, rather than changing the law of any given clonotype.

B.3 The Post-selection Large Deviations Rate Function

To evaluate the infimum in Eq. (14), it is useful to split the set Γ into two non-overlapping subsets Γ_{pre} and Γ_{sel}:

$$\Gamma_{\text{pre}}\stackrel{\text{def}}{=}\Gamma\cap\{\boldsymbol{\nu}:\langle\boldsymbol{\rho},\boldsymbol{\nu}\rangle\leq w_{\text{thy}}\} \tag{B.14}$$

$$\Gamma_{\text{sel}}\stackrel{\text{def}}{=}\Gamma\cap\{\boldsymbol{\nu}:\langle\boldsymbol{\rho},\boldsymbol{\nu}\rangle>w_{\text{thy}}\} \tag{B.15}$$

and to define associated rate functions, as follows:

$$I_{\text{pre}}\stackrel{\text{def}}{=}\inf_{(\boldsymbol{\nu},\boldsymbol{w})\in\Gamma_{\text{pre}}\times\{\boldsymbol{w}:\langle\mathbf{1},\boldsymbol{w}\rangle\leq w_{\text{thy}}\}}[H(\boldsymbol{\nu}|\mu)$$

$$+m\langle\boldsymbol{\pi},\boldsymbol{I}(\boldsymbol{w};\boldsymbol{\nu},\boldsymbol{u},\boldsymbol{\rho})\rangle] \tag{B.16}$$

$$I_{\text{sel}}\stackrel{\text{def}}{=}\inf_{(\boldsymbol{\nu},\boldsymbol{w})\in\Gamma_{\text{sel}}\times\{\boldsymbol{w}:\langle\mathbf{1},\boldsymbol{w}\rangle\leq w_{\text{thy}}\}}[H(\boldsymbol{\nu}|\mu)$$

$$+m\langle\boldsymbol{\pi},\boldsymbol{I}(\boldsymbol{w};\boldsymbol{\nu},\boldsymbol{u},\boldsymbol{\rho})\rangle]. \tag{B.17}$$

For $\boldsymbol{\nu}\in\Gamma_{\text{pre}}$, the infimum is the same as the pre-selection estimate:

$$I_{\text{pre}}=\inf_{\boldsymbol{\nu}\in\Gamma_{\text{pre}}}H(\boldsymbol{\nu}|\mu) \tag{B.18}$$

whereas for $\boldsymbol{\nu}\in\Gamma_{\text{sel}}$, the space of vectors $\boldsymbol{w}$ can be restricted to those for which the TCR signal $\langle\mathbf{1},\boldsymbol{w}\rangle$ *equals* w_{thy}:

$$I_{\text{sel}}=\inf_{(\boldsymbol{\nu},\boldsymbol{w})\in\Gamma_{\text{sel}}\times\{\boldsymbol{w}:\langle\mathbf{1},\boldsymbol{w}\rangle=w_{\text{thy}}\}}[H(\boldsymbol{\nu}|\mu)$$

$$+m\langle\boldsymbol{\pi},\boldsymbol{I}(\boldsymbol{w};\boldsymbol{\nu},\boldsymbol{u},\boldsymbol{\rho})\rangle]. \tag{B.19}$$

Eq. (14), can now be rewritten

$$\lim_{N\to\infty}\frac{1}{N}\ln\mathbb{P}\{\hat{L}\in\Gamma\}=-I_\Gamma\quad\text{where}$$

$$I_\Gamma\stackrel{\text{def}}{=}\min(I_{\text{pre}},I_{\text{sel}})-I_{\text{surv}}. \tag{B.20}$$

Equations (B.18)–(B.20) succinctly represent the statistical structure of the post-selection repertoire. Thus, the post-selection probability of a mature T cell's law $\hat{\boldsymbol{L}}$ belonging to some open set $\Gamma\subset[0,1]^K$ is estimated as $\exp\{-NI_\Gamma\}$. In general, it may not be possible to evaluate $\min(I_{\text{pre}},I_{\text{sel}})$. However, when the elements of Γ all satisfy a constraint of the form $F(\boldsymbol{\nu})=0$, I_Γ can be evaluated using the following auxiliary quantities, which can be determined through the method of Lagrange multipliers:

$$\bar{I}_{\text{pre}}\stackrel{\text{def}}{=}\inf_{\nu\in\Gamma}\boldsymbol{H}(\boldsymbol{\nu}|\mu) \tag{B.21}$$

$$\bar{I}_{\text{sel}}\stackrel{\text{def}}{=}\inf_{(\boldsymbol{\nu},\boldsymbol{w})\in\Gamma\times\{\boldsymbol{w}:\langle\mathbf{1},\boldsymbol{w}\rangle=w_{\text{thy}}\}}[H(\boldsymbol{\nu}|\mu)$$

$$+m\langle\boldsymbol{\pi},\boldsymbol{I}(\boldsymbol{w};\boldsymbol{\nu},\boldsymbol{u},\boldsymbol{\rho})\rangle] \tag{B.22}$$

The required minimum $\min(I_{\text{pre}},I_{\text{sel}})$ is derived from these quantities as follows. If the infimum of Eq. (B.21) is

achieved for some $\boldsymbol{\nu} \in \Gamma_{\text{pre}}$, then $I_{\text{pre}} = \bar{I}_{\text{pre}} \leq \inf_{\boldsymbol{\nu}\in\Gamma_{\text{sel}}} H(\boldsymbol{\nu}|\mu) \leq I_{\text{sel}}$ and thus $\min(I_{\text{pre}}, I_{\text{sel}})L = \bar{I}_{\text{pre}}$. If, on the other hand, $\inf_{\nu\in\Gamma} H(\boldsymbol{\nu}|\mu)$ is achieved for some $\boldsymbol{\nu} \notin \Gamma_{\text{pre}}$, the convexity of the entropy function implies that I_{pre} is achieved on the selection threshold, that is,

$$I_{\text{pre}} = \inf_{\boldsymbol{\nu}\in\Gamma\cap\{\boldsymbol{\nu}:\langle\boldsymbol{\rho},\boldsymbol{\nu}\rangle=w_{\text{thy}}\}} H(\boldsymbol{\nu}|\mu).$$

In this case $\bar{I}_{\text{sel}}$ cannot be achieved for a $\boldsymbol{\nu} \in \Gamma_{\text{pre}}$ which means that $I_{\text{sel}} = \bar{I}_{\text{sel}} \leq I_{\text{pre}}$, whence $\min(I_{\text{pre}}, I_{\text{sel}}) = \bar{I}_{\text{sel}}$.

The term I_{surv} is found as $\min(I_{\text{pre}}, I_{\text{sel}})$ for $\Gamma = (0,1)^K$. Then, when $w_{\text{thy}} > \mu$, $I_{\text{pre}} = \bar{I}_{\text{pre}} = 0$ as the infimum is reached for $\boldsymbol{\nu} = \mu\mathbf{1} \in \Gamma_{pre}$, and thus $I_{\text{surv}} = 0$ as well. When $w_{\text{thy}} < \mu$,

$$I_{\text{surv}} = \bar{I}_{\text{sel}} = \inf_{(\boldsymbol{\nu},\boldsymbol{w})\in(0,1)^K\times\{\boldsymbol{w}:\langle\mathbf{1},\boldsymbol{w}\rangle=w_{\text{thy}}\}} [H(\boldsymbol{\nu}|\mu) + m\langle\boldsymbol{\pi}, \boldsymbol{I}(\boldsymbol{w};\boldsymbol{\nu},\boldsymbol{u},\boldsymbol{\rho})\rangle].$$

For $m = 0$, this gives $I_{\text{surv}} = 0$ (since now $\mu\mathbf{1} \in \Gamma_{\text{sel}}$), while $\lim_{m\to\infty} I_{\text{surv}} = \inf_{\boldsymbol{\nu}\in\{\boldsymbol{\nu}:\langle\boldsymbol{\rho},\boldsymbol{\nu}\rangle = w_{\text{thy}}\}} H(\boldsymbol{\nu}|\mu)$.

When the set Γ is determined by a constraint of the form $\langle\boldsymbol{\alpha},\boldsymbol{\nu}\rangle - \omega = 0$, the rate function $\bar{I}_{\text{sel}}$ is found as

$$\bar{I}_{\text{sel}} = \lambda\omega + m\vartheta w_{\text{thy}} - \sum_{k=1}^{K} \pi_k \ln\{1 - \mu + \mu e_k\} \quad \text{(B.23)}$$

where λ and ϑ are Lagrange multipliers determined by $\langle\boldsymbol{\alpha}, \bar{\boldsymbol{\nu}}(\lambda,\vartheta)\rangle = \omega$ and $\langle\mathbf{1}, \bar{\boldsymbol{w}}(\lambda,\vartheta)\rangle = w_{\text{thy}}$; thus $\bar{I}_{\text{sel}}$ is a function of w_{thy} and ω. Here $\bar{\boldsymbol{\nu}}(\lambda,\vartheta)$ and $\bar{\boldsymbol{w}}(\lambda,\vartheta)$ denote the values attained at the infimum, and depend on the Lagrange multipliers as follows:

$$\bar{\nu}_k = \frac{\mu e_k}{1 - \mu + \mu e_k} \quad \text{and} \qquad \bar{w}_k = \frac{\rho_k \mu e_k x_k}{(1 - \mu + \mu e_k)(1 - u_k + u_k x_k)} \quad \text{(B.24)}$$

where

$$x_k = \exp\{\vartheta\rho_k/(u_k\pi_k)\} \quad \text{and} \quad e_k = \exp\{\lambda\alpha_k/\pi_k + m\ln\{1 - u_k + u_k x_k\}\}. \quad \text{(B.25)}$$

The approximations given in "Statistics of autoreactivity following negative selection" are readily derived by determining the partial derivatives of $\bar{I}_{\text{sel}}$ with respect to w_{thy} and ω.

A convexity argument shows that $I_\Gamma = \bar{I}_{\text{pre}}$ for $\omega \leq \tilde{\omega}$ and $I_\Gamma = \bar{I}_{\text{sel}}$ for $\omega > \tilde{\omega}$, where $\tilde{\omega}$ is determined by the condition $\bar{I}_{\text{pre}} = \bar{I}_{\text{sel}}$, which gives

$$\tilde{\omega} \stackrel{\text{def}}{=} \mu + \sqrt{r_{\boldsymbol{\alpha}}/\varepsilon_{\boldsymbol{\alpha}}}(w_{\text{thy}} - \mu). \quad \text{(B.26)}$$

The minimum of $\bar{I}_{\text{sel}}$ is relevant in the case $w_{\text{thy}} < \mu$ for which it determines I_{surv}. This minimum occurs at $\omega = \hat{\omega}$:

$$\hat{\omega} \stackrel{\text{def}}{=} \mu + \sqrt{r_{\boldsymbol{\alpha}}\varepsilon_{\boldsymbol{\alpha}}}\frac{m}{m + m_{\text{crit}}}(w_{\text{thy}} - \mu). \quad \text{(B.27)}$$

The values $\tilde{\omega}$ and $\hat{\omega}$ determine the critical value w_{crit}, defined in Eq. (21).