Statistical physics of T cell receptor development
and antigen specificity

by

Andrej Košmrlj

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

© Massachusetts Institute of Technology 2011. All rights reserved.

Author

Department of Physics
May 18, 2011

Certified by

Arup K. Chakraborty
Professor of Chemical Engineering, Chemistry and Biological Engineering
Thesis Supervisor

Certified by

Mehran Kardar
Professor of Physics
Thesis Supervisor

Accepted by

Krishna Rajagopal
Professor of Physics
Associate Department Head for Education
Statistical physics of T cell receptor development and antigen specificity

by

Andrej Košmrlj

Submitted to the Department of Physics on May 18, 2011, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physics

Abstract

Higher organisms, such as humans, have an adaptive immune system that usually enables them to successfully combat diverse (and evolving) microbial pathogens. The adaptive immune system is not preprogrammed to respond to prescribed pathogens, yet it mounts pathogen-specific responses against diverse microbes, and establishes memory of past infections (the basis of vaccination). Although major advances have been made in understanding pertinent molecular and cellular phenomena, the mechanistic principles that govern many aspects of an immune response are not known. In this thesis, I illustrate how complementary approaches from the physical and life sciences can help confront this challenge. Specifically, I describe work that brings together statistical mechanics and cell biology to shed light on how key regulators of the adaptive immune system, T cells, are selected to enable pathogen-specific responses. A model of T cell development is introduced and analyzed (computationally and analytically) by employing methods from statistical physics, such as extreme value distributions and Hamiltonian minimization. Results show that selected T cell receptors are enriched in weakly interacting amino acids. Such T cell receptors recognize (i.e. bind sufficiently strongly to) pathogens through several contacts of moderate strength, each of which makes a significant contribution to overall binding. Disrupting any contact by mutating the pathogen is statistically likely to abrogate T cell recognition of the mutated pathogen. We propose that this is the mechanism for the specificity of T cells for unknown pathogens. The T cell development model is also used to discuss one way in which host genetics can influence the selection of T cells and concomitantly the control of HIV infection. A model of the T cell selection process as diffusion in a random field of immobile traps that intermittently turn “on” and “off” is developed to estimate the escape probability of dangerous T cells that could cause autoimmune disease. Finally, and importantly, throughout this thesis, I describe, how the theoretical studies are closely synergistic/complementary with biological experiments and human clinical data.
Thesis Supervisor: Arup K. Chakraborty
Title: Professor of Chemical Engineering, Chemistry and Biological Engineering

Thesis Supervisor: Mehran Kardar
Title: Professor of Physics
To my parents Jože and Veronika and sister Maja.
Acknowledgments

Grad school at MIT turned out to be a very enjoyable experience both inside and outside of school. I feel honored and privileged to have had an opportunity to work with my advisors Arup K. Chakraborty and Mehran Kardar. They provided me with an excellent environment for my studies, introduced me to very interesting and challenging problems, guided me through obstacles, allowed me to attend many interesting scientific conferences, and most importantly taught me a lot of practical things that will help me in the future. I would also like to thank the students and postdocs in our research group for providing a very relaxed work atmosphere.

The biggest role in my life outside of grad school has been the Tang Hall graduate dorm. Participating in intramural sports for Tang introduced me to many new good friends (too many to list them here), helped me to relax from work, and allowed me to try out some new sports (especially american sports) that I would probably never have gotten to learn if I didn’t come here. I also really enjoyed playing soccer with the MIT Graduate Soccer Club.

Finally, I thank my parents Jože and Veronika, and my sister Maja for their love, support and encouragement throughout my life.
Contents

1 Introduction
   1.1 Basic immunology .......................................................... 20
       1.1.1 Two arms of the adaptive immune system ...................... 21
       1.1.2 What is recognition, where does it occur, and what do T cells and B cells recognize? ......................... 22
       1.1.3 T cell development in the thymus gland. ...................... 25
   1.2 Thesis outline ................................................................. 26

2 How the thymus designs antigen-specific and self-tolerant T cell receptor sequences? ......................................................... 29
   2.1 Thymic selection model .................................................... 31
   2.2 Selection against many self-pMHC molecules is required for antigen-specific TCR sequences. .......................... 35
   2.3 Frustration during negative selection strongly constraints selected TCR sequences. .............................. 40
   2.4 Negative selection against many self peptides results in TCR sequences with peptide contact residues enriched in weakly interacting amino acids. 41
   2.5 Antigen specificity is the result of TCR residues binding peptides via multiple moderate interactions. ............. 44
   2.6 Escape of T cells, which could trigger an autoimmune disease .... 48
   2.7 Conclusions ........................................................................ 51

3 Thymic selection of T-cell receptors as an extreme value problem 55
3.1 Thymic selection model ........................................ 56
  3.1.1 Extreme value problem .................................... 58
  3.1.2 Hamiltonian minimization ................................. 60
3.2 T cell recognition of foreign peptides ......................... 63
3.3 Conclusions ..................................................... 73

4 Effects of thymic selection of the T-cell repertoire on HLA class
  I-associated control of HIV infection ......................... 75
  4.1 HLA-B57 binds fewer types of self peptides ................. 76
  4.2 Thymic selection against fewer types of self peptides results in a more
      cross-reactive T-cell repertoire .............................. 77
  4.3 Cross-reactivity of T cell repertoire correlates with the control of HIV
      infection ....................................................... 81
  4.4 Conclusions ..................................................... 87

5 Thymic selection of T cells as diffusion with intermittent traps .... 89
  5.1 Classical trapping problem ................................... 90
  5.2 Intermittent traps ............................................ 95
  5.3 Escape probability of autoimmune T cells .................. 99

6 Conclusions ..................................................... 103

A Robustness of the thymic selection model ....................... 107
  A.1 Robustness to changes in interaction potential ............ 107
  A.2 Robustness to changes in other parameters ................ 111

B Accuracy of prediction algorithms for peptide binding affinities to
  HLA and Mamu alleles ............................................ 119
  B.1 Measures of accuracy of prediction algorithms ............. 120

C Host–pathogen interaction dynamics ................................ 129
  C.1 Host–pathogen interaction dynamics model ................ 129
  C.2 Host–pathogen interaction dynamics for simplified model .... 134
C.3 Robustness of the effects of cross-reactivity to variations in parameters and models. ......................... 136
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Lymphocyte recognition of signatures of pathogens.</td>
<td>23</td>
</tr>
<tr>
<td>1-2</td>
<td>Specificity/degeracy of TCR recognition of foreign peptides.</td>
<td>24</td>
</tr>
<tr>
<td>1-3</td>
<td>Thymic selection of T cells.</td>
<td>26</td>
</tr>
<tr>
<td>2-1</td>
<td>Schematic representation of the interface between TCR and pMHC complexes.</td>
<td>32</td>
</tr>
<tr>
<td>2-2</td>
<td>Cartoon representation of the three regimes of values of TCR-MHC interactions ( (E_c) ).</td>
<td>37</td>
</tr>
<tr>
<td>2-3</td>
<td>Selection against many peptides in the thymus results in a larger number of important contacts characterizing antigen recognition.</td>
<td>39</td>
</tr>
<tr>
<td>2-4</td>
<td>Schematic description of frustration due to negative selection.</td>
<td>41</td>
</tr>
<tr>
<td>2-5</td>
<td>Selection against many types of peptides in the thymus results in selected TCRs with peptide contact residues with an enhanced frequency of amino acids that interact weakly with all other amino acids.</td>
<td>43</td>
</tr>
<tr>
<td>2-6</td>
<td>Distribution of amino acid-amino acid contact energies characterizing interactions between selected reactive TCRs and antigenic peptides suggest the basis for specificity.</td>
<td>45</td>
</tr>
<tr>
<td>2-7</td>
<td>Amino acid frequencies of recognized antigenic peptides are enriched with strongly interacting amino acids.</td>
<td>46</td>
</tr>
<tr>
<td>2-8</td>
<td>Mechanism for specificity and degeneracy of TCR recognition of antigenic peptides.</td>
<td>47</td>
</tr>
<tr>
<td>2-9</td>
<td>Ratio of the numbers of escaped autoimmune TCRs and selected TCRs.</td>
<td>50</td>
</tr>
</tbody>
</table>
2-10 A bar code scanning model for specificity of TCR recognition of antigenic peptides. ......................................................... 53

3-1 Schematic presentation of the choice of parameter $\beta$ with entropy maximization. .................................................. 61

3-2 Phase diagram of the parameter $\beta$. .................................................. 62

3-3 Amino-acid composition of selected TCR sequences. .................................................. 64

3-4 Phase diagram of the parameter $\beta^*$. .................................................. 66

3-5 Amino-acid composition of recognized foreign peptides. .................................................. 67

3-6 Amino-acid composition of foreign peptides recognized by uncorrelated TCRs. .................................................. 69

3-7 Amino-acid composition of foreign peptides recognized by a complete pool of selected TCRs. .................................................. 72

4-1 Thymic selection against fewer self peptides leads to a more cross-reactive T-cell repertoire. .................................................. 80

4-2 Dynamic model of host–pathogen interactions. .................................................. 82

4-3 Model shows reduced set-point viral load for people with more cross-reactive T-cell repertoire. .................................................. 83

4-4 Virus diversity and immune pressure dependence on the cross-reactivity of T cell repertoire. .................................................. 84

4-5 HLA-B alleles associated with greater ability to control HIV correlate with smaller self-peptide binding propensities. .................................................. 86

5-1 Collapse of survival probabilities. .................................................. 94

5-2 Model of a random walker in a random field of intermittent traps. .................................................. 95

5-3 Survival probabilities for a random walker in a random field of intermittent traps. .................................................. 97

5-4 Caging effect. .................................................. 98

A-1 Amino acid frequencies in selected TCRs for the MJ matrix. .................................................. 108
A-2 Amino acid frequencies in selected TCRs for the semi-random matrix 
\((J + R)/2\). .............................. 109
A-3 Amino acid frequencies in selected TCRs for the random matrix \(R\). . 110
A-4 Thymic selection results for more complex interactions between a TCR 
and a pMHC. .............................. 112
A-5 Thymic selection using soft thresholds for positive and negative selection.113
A-6 Thymic selection using amino acid frequencies from mouse proteome. 114
A-7 Thymic selection results for different TCR binding strengths to MHC. 115
A-8 Results of thymic selection as a function of the number of self peptides 
\(M\) in the thymus. .............................. 116
A-9 Probabilities of TCR selection. .............................. 117
A-10 Frequency distribution of amino acids in TCR that are in contact with 
peptide. .............................. 118

B-1 Accuracy of predictive peptide binding algorithms for HLA-B*5701 
and HLA-B*2705. .............................. 126
B-2 Accuracy of predictive peptide binding algorithms for HLA-B*0801. . 127
B-3 Accuracy of predictive peptide binding algorithms for HLA-B*3501. . 128

C-1 Random energy-like model for generating the matrix describing recogni-
tion of pMHCs by CD8\(^+\) T cells. .............................. 133
C-2 Simplified model for host–pathogen interaction dynamics. ............. 135
C-3 Simplified model shows a reduced set-point viral load for people with 
a more cross-reactive T-cell repertoire. .............................. 137
C-4 Virus diversity and immune pressure dependence on the cross-reactivity 
of T cell repertoire for the simplified host–pathogen model. ............. 138
C-5 Anticorrelation of simulated peak viral loads with percent contribution 
of the dominant epitope to the total CTL response for the simplified 
host–pathogen model. .............................. 139
C-6 Peak viral load versus average number of CD8\(^+\) clones recognizing a 
pMHC in each simulation in the simplified host–pathogen model. . . 140
C-7  Insensitivity of qualitative results to changes in CD8+ T cell activation rate for the simplified host-pathogen model. .......................... 141
C-8  Anticorrelation of simulated viral loads with percent contribution of the dominant epitope to the total CTL response. ................. 142
C-9  Peak viral load versus average number of CD8+ clones recognizing a pMHC in each simulation. .............................. 143
C-10 Insensitivity of qualitative results to changes in CD8+ T cell activation rate. ................................................................. 144
C-11 Setpoint viral load versus percent contribution of immunodominant epitope for different rates of peptide presentation. .......... 145
List of Tables

1.1 Glossary of immunological terms.

2.1 Amino acid frequencies of *Homo sapiens*, mouse and *Listeria monocytogenes* proteomes.

4.1 Alleles with significant association of HIV control or progression.

B.1 Accuracy of predictive algorithms for 9-mer peptide binding by HLA-B and Mamu alleles.

C.1 Parameters of host-pathogen interaction model.

C.2 Parameters of simplified host-pathogen interaction model.
Chapter 1

Introduction

The immune system of an organism combats invading pathogens, thereby protecting the host from disease. Jawed vertebrates, such as humans, have an adaptive immune system that enables them to mount pathogen-specific immune responses (Janeway et al., 2004). The importance of the adaptive immune response for human health is highlighted by the opportunistic infections that afflict individuals with compromised adaptive immune systems [e.g., those who have progressed to AIDS after infection with the human immunodeficiency virus (HIV)]. The flexible adaptive immune system can also go awry, and many diseases (e.g., multiple sclerosis and type I diabetes) are the consequence of the adaptive immune system failing to discriminate between markers of self and nonself. The suffering caused by autoimmune diseases and the need to combat diverse infectious agents have motivated a great deal of experimental research aimed toward understanding how the adaptive immune system is regulated. These efforts have led to some spectacular discoveries (Allison et al., 1982; Chan et al., 1992; Dialynas et al., 1983; Gallegos and Bevan, 2006; Hedrick et al., 1984; Hogquist et al., 2005; Irving and Weiss, 1991; Unanue, 1984; Yanagi et al., 1984). Yet an understanding of the principles that govern the emergence of an immune or autoimmune response has proven to be elusive. The practical impact of this missing knowledge is highlighted by the inability to design vaccines against many scourges on the planet (e.g., HIV).

This chapter is based on the publication (Chakraborty and Košmrlj, 2010).
An important barrier for the development of mechanistic principles that describe adaptive immunity is that the pertinent processes involve cooperative dynamic events with many participating components that must act collectively for an immune or autoimmune response to emerge. Moreover, these processes span a spectrum of timescales and length scales that range from interactions between molecules in cells to phenomena that affect the entire organism; feedback loops between processes on different spatiotemporal scales are also important. It is often hard to intuit underlying principles from experimental observations because of the complexity of these hierarchically organized collective processes. The importance of stochastic effects further confounds intuition.

Statistical mechanics can relate observations to the underlying microscopic stochastic events that occur in a complex interacting system. Statistical mechanical theory, associated computations, and complementary experiments have therefore helped uncover mechanisms underlying complex physical and chemical phenomena. In this thesis I describe work that brings together statistical mechanics and cell biology to uncover new concepts in immunology. To properly define the problem, I begin with a brief exposition of basic immunology.

1.1 Basic immunology

Higher organisms are constantly exposed to infectious microbial pathogens. Yet the development of infectious diseases is relatively rare. This is because diverse types of cells that compose the innate immune system are efficient in preventing pathogenic microorganisms from establishing an infection. The components of the innate immune system respond to common features of diverse pathogens, but are not specific for individual pathogens. Some bacteria and many viruses can evade or overcome the innate mechanisms of host defense. The adaptive immune system mounts pathogen-specific immune responses against such invading microorganisms. Adaptive immunity also establishes memory of past infections, thereby conferring the ability to mount rapid immune responses to pathogens encountered previously. This immunological
memory is the basis for vaccination.

1.1.1 Two arms of the adaptive immune system

The adaptive immune system has two arms, called cellular and humoral immunity. T lymphocytes (T cells) and B lymphocytes (B cells) are the key regulators of cellular and humoral immunity, respectively. T cells and B cells express immunoglobulin proteins on their surfaces, which are called T cell receptors (TCRs) and B cell receptors (BCRs), respectively. The genes encoding these receptors are inherited as gene segments that stochastically recombine during the synthesis of T cells and B cells in the bone marrow. Each gene assembled in a given lymphocyte is thus distinct, enabling the generation of a great diversity of T cells and B cells expressing different receptors. Different lymphocytes can potentially respond to specific pathogens as distinct receptors can potentially recognize (i.e., bind sufficiently strongly to) molecular signatures of specific foreign invaders. Thus, the adaptive immune system can mount pathogen-specific responses against diverse infectious microbes.

The diverse lymphocytes bearing different TCRs and BCRs that are generated in the bone marrow by the stochastic recombination of gene segments do not all become part of an organism’s army of T cells and B cells that battle pathogens. Rather, T cells and B cells undergo development processes that allow only a small fraction of the generated cells to become part of an organism’s repertoire of lymphocytes. T cells develop in an organ called the thymus (the T stands for thymus). B cells develop in the bone marrow (the B stands for bone marrow) and also, upon activation, in lymphoid organs. The focus of this thesis is on studies aimed toward understanding how developmental processes in the thymus shape the T cell repertoire such that they exhibit both remarkable pathogen specificity and the ability to combat myriad pathogens. To consider these issues, we note a few more biological facts.
1.1.2 What is recognition, where does it occur, and what do T cells and B cells recognize?

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

If a lymphocyte recognizes pathogens in a lymph node, a series of intracellular biochemical reactions occurs (called signaling) that results in gene transcription programs that cause the lymphocyte to become activated; i.e., it begins to proliferate and acquire the ability to carry out functions that can mediate an immune response. Activated lymphocytes thus generated, bearing receptors specific for the infecting pathogen, then leave the lymph node and enter the blood via lymphatic vessels. When they encounter the same pathogen’s molecular markers in the blood or tissues, they can carry out effector functions to eliminate the infection. For example, certain kinds of activated T cells can kill cells infected by the pathogen, thereby killing the pathogen as well.

The BCRs and TCRs expressed on B cells and T cells can bind to species that are called ligands (Figure 1-1). B cells protect against pathogens in blood or extracellular spaces. The ligands of the BCR include proteins, fragments of proteins, and molecules on the surface of viruses or bacteria. T cells evolved to combat intracellular pathogens. Proteins synthesized by intracellular pathogens are cut up into short peptide fragments by enzymes in cells harboring the pathogen. These peptide fragments can potentially bind to protein products of the host’s major histocompatibility complex (MHC) genes. There are two kinds of MHC proteins, called MHC class I and
Figure 1-1: Lymphocyte recognition of signatures of pathogens. (a) T cells (CTL—cytotoxic T lymphocyte.). Antigen-presenting cells (APCs) engulf pathogens and process their proteins into short peptides, which are bound to major histocompatibility complex (MHC) proteins and presented on the surface. T cell receptors (TCRs) bind to peptide MHCs, and sufficiently strong binding enables intracellular signaling and gene transcription, leading to T cell activation. APCs also present self-peptides derived from self-proteins, but typically T cells are not activated by them. (b) B cells. B cell receptors (BCRs) bind directly to antigens and their products, and sufficiently strong binding results in productive intracellular signaling and internalization of the antigen, bound to BCR. The antigen is processed by the B cell, which then presents the corresponding peptide MHCs on its surface. Recognition of these peptide MHCs by an activated T helper cell’s TCR is usually necessary for B cell activation. This figure is adapted from Figure 1 in (Chakraborty and Košmrlj, 2010).

MHC class II. Typically, a human will have up to six types of MHC class I proteins, and up to six types of MHC class II proteins. Pathogen-derived peptides (p) bound to MHC proteins are ultimately expressed on the surface of APCs and infected cells. These pMHCs are the TCR ligands.

When we say that a T cell recognizes a particular pathogen-derived pMHC, what we mean is that its TCR binds to it sufficiently strongly, which allows productive intracellular signaling and activation to occur. T cells activated by peptides presented by MHC class II proteins proliferate and differentiate into many cell types called T helper cells (for reasons noted below). T cells activated by peptides presented by MHC class I molecules become cytotoxic T lymphocytes (CTLs). When activated
TCR recognition of pathogen-derived pMHC molecules is both highly specific and degenerate. It is specific because if a TCR recognizes a peptide (green), most point mutations of the peptide's amino acids (red) abrogate recognition. However, a given TCR can also recognize diverse peptides (green, blue, yellow).

CTLs encounter cells in tissues that express the pMHC molecules that originally activated them, they can kill these cells by secreting various chemicals.

TCR recognition of pathogen-derived pMHC molecules is both highly specific and degenerate (Figure 1-2). It is specific because if a TCR recognizes a pMHC molecule, most point mutations of the peptide's amino acids abrogate recognition (Huseby et al., 2005, 2006). However, a given TCR can also recognize diverse peptides (Eisen and Chakraborty, 2010; Hemmer et al., 1998; Kersh and Allen, 1996; Misko et al., 1999; Sloan-Lancaster and Allen, 1996; Unanue, 1984). This specificity-degeneracy conundrum is made vivid by dividing the world of peptides into classes, with the members of each class having sequences that are closely related. For example, peptides within a class could differ by just point mutations. A TCR can discriminate quite well between peptides within a class of closely related peptide sequences (as point mutants of the peptides it recognizes are not recognized with high probability). But, at the same time, TCRs are not so good at distinguishing between peptides with different sequences as a given TCR can recognize many peptides. Understanding how the TCR repertoire obtains these properties is the main focus of Chapter 2.

Similarly, when we say that a B cell recognizes a ligand, we mean that its BCR
binds sufficiently strongly to it, which results in signaling and also causes the internalization of the pathogen. pMHCs, with the peptide derived from the internalized pathogen, are presented on the B cell surface. Signaling induced by the binding of these class II pMHCs with a T helper cell’s TCR (activated by the same pMHC) augments BCR signaling to activate B cells. Activated B cells proliferate and differentiate into plasma cells that secrete a soluble form of its BCR. These soluble immunoglobulins are called antibodies. Antibodies act on pathogens in extracellular spaces and in blood in a variety of different ways to help clear infections. Both T and B cells also differentiate into memory cells that mount rapid immune responses upon reinfection with the same pathogen.

The lymphocyte signaling network does not exhibit a continuous increase in response as the stimulus (e.g., TCR-pMHC binding strength) is progressively increased. Rather, it only responds strongly above a threshold stimulus level. This feature is necessary for sharp discrimination between recognized and unrecognised ligands.

1.1.3 T cell development in the thymus gland.

After synthesis in the bone marrow, immature T cells (thymocytes) go to the thymus (Figure 1-3; (Hogquist et al., 2005; Siggs et al., 2006; Starr et al., 2003; von Boehmer et al., 2003; Werlen et al., 2003)), an organ located behind the sternum. Thymocytes interact with a variety of self-pMHC molecules expressed on the surface of thymic epithelial cells as well as hematopoietically derived macrophages and dendritic cells, where self peptides are derived from diverse parts of the host proteome. For a T cell to exit the thymus and become part of the host’s repertoire of T cells, it must pass the following two tests: (a) It must not be negatively selected. That is, its TCR must not bind to any self-pMHC molecule with a binding free energy that exceeds a threshold for negative selection. Such strong binding is likely to cause deletion of the T cells. (b) It must bind at least one self-pMHC molecule with a binding free energy that exceeds another threshold for positive selection. The negative selection process aims to delete dangerous T cells, that could be activated by self pMHCs and cause an autoimmune disease. The positive selection process ensures that TCRs of
Figure 1-3: Immature T cells (thymocytes) develop in the thymus. Thymocytes migrate through the thymus and interact with diverse self peptide major histocompatibility complexes (self-pMHCs) presented on the surface of thymic antigen presenting cells (APCs). A T cell’s receptor (TCR) must bind to at least one of these self-pMHCs weakly to exit the thymus and become a part of the individual’s T cell repertoire (positive selection). A T cell with a TCR that binds to any self-pMHC with an affinity that exceeds a sharply defined threshold dies in the thymus (negative selection). This figure is adapted from Figure 3a in (Chakraborty and Košmrlj, 2010).

selected T cells bind to MHC – this is called MHC restriction. Positively selected T cells become part of an organism’s army that battle pathogens if they can avoid negative selection.

### 1.2 Thesis outline

In this thesis, I present studies that enable an understanding of how the developmental processes in the thymus shape the T cell repertoire such that adaptive immunity exhibits both remarkable pathogen specificity and the ability to combat myriad pathogens.

Chapter 2 sheds light on the specificity and degeneracy puzzle of the TCR recognition of foreign peptides (Figure 1-2). In this chapter we develop a statistical mechanical model of the thymic selection process of T cells. Computational studies of this model are used to characterize the properties of the selected T cell repertoire, which are then used to describe the mechanism behind their specificity.

In Chapter 3, the model of thymic selection developed in Chapter 2 is analyzed analytically by employing methods from statistical physics, such as extreme value distributions and Hamiltonian minimization.
Chapter 4 investigates how some people are able to control the HIV infection even without therapy. Genetic studies show that certain MHC types appear more frequently in these people. This chapter discusses how some of these MHC types may affect thymic selection, such that the statistical properties of the selected T cell repertoire's characteristics provide one contributing factor for more efficient control of HIV infection.

Chapter 5 was motivated by a question of the escape probability of dangerous T cells from the thymus that can be activated by self peptides. Mimicking the thymic selection, this chapter studies a problem of diffusion in a random field of immobile traps that intermittently switch “on” and “off”. The results of this study are used to estimate the escape probability of dangerous T cells that can be activated by self peptides.

The introduction and other chapters are based on published papers (Chakraborty and Košmrlj, 2010; Košmrlj et al., 2008, 2009, 2010; Košmrlj, 2011).

For the benefit of physicists not familiar with immunology, we present here a glossary of immunological terms used throughout the thesis (Table 1.1).
Table 1.1: Glossary of immunological terms.

<table>
<thead>
<tr>
<th>term</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>antigen</td>
<td>a foreign substance that induces an immune response in the body</td>
</tr>
<tr>
<td>pathogen</td>
<td>a bacterium, virus, or other microorganism that can cause disease</td>
</tr>
<tr>
<td>APC</td>
<td>antigen-presenting cell. They express molecular signatures (peptides) of the ingested antigens on their surface.</td>
</tr>
<tr>
<td>peptide</td>
<td>a short piece of protein presented to T cell receptors in a groove of MHC proteins</td>
</tr>
<tr>
<td>HIV epitope</td>
<td>a peptide derived from HIV proteins</td>
</tr>
<tr>
<td>immunodominant epitope</td>
<td>epitope that elicits the most significant response from the adaptive immune system</td>
</tr>
<tr>
<td>MHC</td>
<td>major histocompatibility complex. MHC proteins present peptides to T cell receptors and are classified into classes I and II.</td>
</tr>
<tr>
<td>HLA</td>
<td>human MHC proteins</td>
</tr>
<tr>
<td>pMHC</td>
<td>peptide-MHC complex</td>
</tr>
<tr>
<td>TCR</td>
<td>T cell receptor. T cell receptors bind to pMHCs.</td>
</tr>
<tr>
<td>CDR loops</td>
<td>TCR consists of 3 CDR loops. Highly variable CDR3 loop binds to peptides and more conserved CDR1 and CDR2 loops bind to MHC.</td>
</tr>
<tr>
<td>CD8+ T cell</td>
<td>T cell receptors of those T cells bind to MHC class I proteins</td>
</tr>
<tr>
<td>CD4+ T cell</td>
<td>T cell receptors of those T cells bind to MHC class II proteins</td>
</tr>
<tr>
<td>CTL</td>
<td>cytotoxic T lymphocytes is CD8+ T cell. Activated CTLs kill infected cells by secreting various chemicals.</td>
</tr>
<tr>
<td>T helper cell</td>
<td>CD4+ T cell. T helper cells help activate B cells.</td>
</tr>
<tr>
<td>thymus gland</td>
<td>An organ, where immature T cells are screened against self peptide-MHCs.</td>
</tr>
<tr>
<td>thymocyte</td>
<td>immature T cell that goes to the thymus</td>
</tr>
<tr>
<td>autoimmune disease</td>
<td>a disease caused by antibodies or lymphocytes produced against substances naturally present in the body</td>
</tr>
<tr>
<td>alloreactivity</td>
<td>The reaction between the immune cells of one individual and the cells of another individual, e.g., the T cell response to non-self MHC molecules</td>
</tr>
</tbody>
</table>
Chapter 2

How the thymus designs antigen-specific and self-tolerant T cell receptor sequences?

T lymphocytes (T cells) orchestrate adaptive immune responses that clear pathogens from infected hosts. T cells recognize short peptides (p) derived from antigenic proteins bound to protein products of the MHC genes. Recognition occurs when T cell receptor (TCR) proteins expressed on T cells bind sufficiently strongly to antigen-derived pMHC complexes on the surface of antigen-presenting cells (Figure 1-1a). A diverse repertoire of self-pMHC–tolerant TCR sequences is shaped during development of T cells in the thymus by processes called positive and negative selection. In this Chapter we combine computational models and analysis of experimental data to parse the contributions of positive and negative selection to the design of TCR sequences that recognize antigenic peptides with specificity, yet also exhibit cross-reactivity. A dominant role for negative selection in mediating antigen specificity of mature T cells and a molecular mechanism for TCR recognition of antigen are described.

Because T cell receptor (TCR) genes undergo stochastic somatic rearrangement, most T cells express a distinct TCR, thereby enabling the T cell population to recog-

This chapter is based on the publication (Kosmrlj et al., 2008).
nize many different antigenic short peptide (p)MHC complexes. TCR recognition of pMHC is both specific and degenerate (Figure 1-2). It is specific, because if a TCR recognizes a particular pMHC complex, most mutations to the peptide amino acids abrogate recognition (Huseby et al., 2005, 2006). It is degenerate because a given TCR can interact productively with several antigenic peptides (Hemmer et al., 1998; Kersh and Allen, 1996; Misko et al., 1999; Sloan-Lancaster and Allen, 1996; Unanue, 1984). pMHC complexes where the peptide is derived from the cell’s own proteins are also displayed on antigen-presenting cell (APC) surfaces. TCRs are self-tolerant because they bind weakly to these “self”-pMHC complexes, thereby avoiding frequent autoimmune responses.

The diverse, specific/degenerate, and self-tolerant T cell repertoire is designed during T cell development in the thymus (Hogquist et al., 2005; Jameson et al., 1995; Siggs et al., 2006; von Boehmer et al., 2003; Werlen et al., 2003). Immature T cells (thymocytes) interact with a variety of self-pMHC molecules expressed on the surface of thymic epithelial cells as well as hematopoietically derived macrophages and dendritic cells. Thymocytes expressing a TCR that binds with high affinity to any self-pMHC molecule are deleted in the thymus (a process called negative selection). However, a thymocyte’s TCR must also bind sufficiently strongly to at least one type of self pMHC complex to receive survival signals and emigrate from the thymus (a process called positive selection).

Signaling events, gene transcription programs, and cell migration during T cell development in the thymus have been studied extensively (Borghans et al., 2003; Bousso et al., 2002; Daniels et al., 2006; Detours et al., 1999; Detours and Perelson, 1999; Hogquist et al., 2005; Jameson et al., 1995; Scherer et al., 2004; Siggs et al., 2006; von Boehmer et al., 2003; Werlen et al., 2003). Despite important advances, how interactions with self-pMHC complexes in the thymus shape the peptide-binding properties of selected TCR amino acid sequences such that mature T cells exhibit their special properties is poorly understood.

Recent experiments carried out by Huseby et al. (2005, 2006) provided important clues in this regard. These experiments determined differences in how T cells interact
with foreign (antigenic) pMHC depending on whether they developed in conventional mice that display a diverse array of self-pMHC complexes in the thymus or if they develop in mice that were engineered to express only one type of peptide in the thymus. For T cells that develop in conventional mice, T cell recognition of antigenic pMHC was found to be sensitive to most mutations of the antigenic peptide's amino acids. In contrast, T cells selected in mice with only one type of peptide in the thymus were much more peptide-degenerate, with some T cells being tolerant to most mutations of antigenic peptide amino acids.

We reasoned that a detailed understanding of the origin of these experimental results may shed light on the broader question of how the thymus designs diverse self-tolerant TCR sequences that mediate specific/degenerate antigen recognition. Toward this end, we studied a computational model of thymic selection.

2.1 Thymic selection model

To describe the interactions between TCRs and pMHC complexes, we represent them as strings of sites (Figure 2-1). Each site on a TCR can interact with the corresponding site on a pMHC molecule. Such string models for studying TCR–pMHC interactions have been used to study various issues, including thymic selection (Chao et al., 2005; Detours et al., 1999; Detours and Perelson, 1999), and employed simplified representations of amino acids (e.g., a string of numbers, bits, etc.). From the standpoint of our work, the most pertinent result revealed by these past studies are calculations showing that negative selection reduces TCR cross-reactivity. The mechanistic reasons underlying this numerical result or how it relates to amino acid sequences of selected TCRs were not described. The goal of this Chapter is to elucidate how the diversity of endogenous peptides bound to host MHC proteins encountered in the thymus determines the amino acid sequences of peptide contact residues on selected TCRs and how such TCRs are antigen specific while also being cross-reactive and self-tolerant.

The specific features of our model were chosen to address these issues and to
Figure 2-1: Schematic representation of the interface between TCR and pMHC complexes. The CDR3 loop segment of TCR that is in contact with peptides is highly variable and modeled by a string of \( N \) amino acids. The peptide is also modeled by a sequence of length \( N \), and the binding free energy is computed as a sum of pairwise interactions. We do not explicitly consider TCR sites (CDR1 and CDR2 loops) in contact with MHC, as they are more or less conserved, and only assign them a net interaction free energy \( E_c \). This figure is adapted from Figure 3b in (Chakraborty and Kosmrlj, 2010).

relate the results closely to known experimental data such as that of Huseby et al. (2005, 2006). Because Huseby et al. used transgenic mice that expressed a single type of MHC, we divided the string of sites on the pMHC molecule into a conserved part representing the MHC and a variable part representing the peptides. One could also view the variable sites more generally as representative of the peptides and the variable residues of the MHC. The CDR1 and CDR2 loops of the TCR mostly contact MHC residues, whereas the CDR3 loop primarily contacts the peptide residues. We partitioned the TCR interaction sites into two parts: a region representing the CDR1 and CDR2 loops and a part that mimics the CDR3 loop. Because the CDR3 loops are hypervariable, the amino acids of the peptide contact residues of the CDR3 region are explicitly considered, whereas those of the less variable CDR1 and CDR2 regions are not (Figure 2-1). For ease of reference, the CDR3 sites are called, variable. These variable sites represent only those CDR3 amino acids that contact peptide amino acids (or variable MHC residues). Thus, we do not explicitly treat the conformation
Table 2.1: Amino acid frequencies of *Homo sapiens*, mouse and *Listeria monocytogenes* proteomes.

<table>
<thead>
<tr>
<th>amino acid</th>
<th>Homo sapiens</th>
<th>Mus musculus (house mouse)</th>
<th>Listeria monocytogenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0692</td>
<td>0.0681</td>
<td>0.0774</td>
</tr>
<tr>
<td>C</td>
<td>0.0225</td>
<td>0.0228</td>
<td>0.0061</td>
</tr>
<tr>
<td>D</td>
<td>0.0476</td>
<td>0.0481</td>
<td>0.0544</td>
</tr>
<tr>
<td>E</td>
<td>0.0718</td>
<td>0.0700</td>
<td>0.0744</td>
</tr>
<tr>
<td>F</td>
<td>0.0359</td>
<td>0.0369</td>
<td>0.0453</td>
</tr>
<tr>
<td>G</td>
<td>0.0658</td>
<td>0.0641</td>
<td>0.0667</td>
</tr>
<tr>
<td>H</td>
<td>0.0261</td>
<td>0.0263</td>
<td>0.0178</td>
</tr>
<tr>
<td>I</td>
<td>0.0434</td>
<td>0.0439</td>
<td>0.0784</td>
</tr>
<tr>
<td>K</td>
<td>0.0576</td>
<td>0.0576</td>
<td>0.0716</td>
</tr>
<tr>
<td>L</td>
<td>0.0985</td>
<td>0.0993</td>
<td>0.0951</td>
</tr>
<tr>
<td>M</td>
<td>0.0215</td>
<td>0.0221</td>
<td>0.0275</td>
</tr>
<tr>
<td>N</td>
<td>0.0360</td>
<td>0.0358</td>
<td>0.0462</td>
</tr>
<tr>
<td>P</td>
<td>0.0636</td>
<td>0.0619</td>
<td>0.0347</td>
</tr>
<tr>
<td>Q</td>
<td>0.0481</td>
<td>0.0479</td>
<td>0.0346</td>
</tr>
<tr>
<td>R</td>
<td>0.0568</td>
<td>0.0563</td>
<td>0.0365</td>
</tr>
<tr>
<td>S</td>
<td>0.0836</td>
<td>0.0850</td>
<td>0.0580</td>
</tr>
<tr>
<td>T</td>
<td>0.0536</td>
<td>0.0541</td>
<td>0.0611</td>
</tr>
<tr>
<td>V</td>
<td>0.0598</td>
<td>0.0609</td>
<td>0.0704</td>
</tr>
<tr>
<td>W</td>
<td>0.0123</td>
<td>0.0120</td>
<td>0.0093</td>
</tr>
<tr>
<td>Y</td>
<td>0.0263</td>
<td>0.0269</td>
<td>0.0345</td>
</tr>
</tbody>
</table>

of the CDR3 loop, which would be necessary if the entire sequence of CDR3 amino acids was considered. Similarly, because peptides bound to MHC are short, peptide conformation is not an important variable. Although we vary the peptide length (data not shown), most results we present in this Chapter are for peptides that are 10 amino acids long\(^1\).

We generate panels of TCR and self pMHC molecules on the computer by picking amino acids for the peptides and peptide contact residues on the CDR3 loops of the TCR according to the probabilities with which amino acids appear in the human (or mouse) proteome (Table 2.1 and (Flicek et al., 2008; Košmrlj et al., 2008)). Antigenic

\(^1\) The number of contacts between a TCR and a peptide is ~5 and in next Chapters we are using sequences of length 5 amino acids. The qualitative results presented in this chapter do not depend on the choice of the sequence length (Chapter 3).
peptides are generated using the frequency of occurrence of amino acids in *Listeria monocytogenes*, a common bacterial pathogen (Table 2.1 and (Kosmrlj et al., 2008; Moszer et al., 1995)). To assess the effects of thymic selection as well as antigen recognition, we evaluate the free energy of interaction between TCR-pMHC pairs. The interaction free energy between the CDR1 and CDR2 regions of TCRs and the MHC is given a value equal to $E_c$ (and it is varied to describe different TCRs). The total interaction free energy equals the sum of $E_c$ and the value obtained by aligning the TCR and pMHC amino acids that are treated explicitly by adding the pairwise interactions between corresponding amino acids. For a given TCR–pMHC pair, the total interaction free energy is

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

(2.1)

where $E_c$ is defined above, and $J(t_i, s_i)$ is the interaction free energy between the $i$th amino acids on the variable part of the TCR ($t_i$) and the peptide ($s_i$), respectively, and $N$ is the length of the variable regions. The matrix $J$ encodes the values of interaction energies between specific types of amino acids. For most results presented, $J$ was taken to be the parameterized potential due to Miyazawa and Jernigan (MJ matrix) which has been used fruitfully to study proteins (Li et al., 1997; Miyazawa and Jernigan, 1996). However, we also used other potentials (*vide infra*), including ones where the interaction between a pair of juxtaposed amino acids depends on the neighboring residues, to show that our qualitative results and mechanistic insights are independent of this choice (see Figures A-1 – A-4). We express free energy values in units of the thermal energy, $k_B T$, where $k_B$ is Boltzmann’s constant, and $T = 37^\circ$ C is room temperature, at which the thermal energy equals 0.6 kcal mol$^{-1}$. We emphasize that the purpose of our study is not to compute specific values of energies but to use them to obtain qualitative mechanistic insights.

Recent experiments show that negative selection occurs when the TCR–pMHC interaction affinity exceeds a sharply defined threshold (Daniels et al., 2006). Because affinity correlates directly with the free energy gained upon binding, in our model, if
the interaction free energy, $E_{\text{int}}$, between a TCR and self-pMHC is more attractive than (exceeds) a threshold value ($|E_{\text{int}}| > |E_n|$), this TCR is negatively selected. It is possible that the off-rate characterizing TCR–pMHC binding, rather than affinity, determines ligand potency, and, indeed, ligands that induce positive and negative selection are separated by a sharp boundary in off-rate as well. Off-rate correlates with the free-energy barrier associated with dissociation of the TCR–pMHC complex. For a related set of reactions, this barrier and the binding free energy scale similarly (Edwards, 1954; Swain and Scott, 1953) and so use of the interaction free energy should correlate with trends in off-rate as well. The ability of a pMHC ligand to stimulate positive selection does not go to zero abruptly (Daniels et al., 2006). In our model, if the interaction free energy between a particular TCR–pMHC pair exceeds a threshold value, $E_p$, the TCR is positively selected. Replacing the soft threshold associated with positive selection with a sharp boundary does not affect qualitative results (Figure A-5) because we find that the characteristics of peptide binding residues on selected T cells are largely shaped by negative selection. The effects of varying $E_p$ and $E_n$ over wide ranges are described in the context of our results.

2.2 Selection against many self-pMHC molecules is required for antigen-specific TCR sequences.

We first tested whether our computational model could recapitulate the experimental observation (Huseby et al., 2005, 2006) that T cell recognition of an antigenic peptide is sensitive to mutations at many peptide sites for T cells selected against many endogenous thymic peptides, whereas very few sites on the antigenic peptide are important for recognition for T cells selected in mice that express one type of peptide

---

2 The relationship between the change in free energy at equilibrium (related to affinity) and the free energy barrier for the reaction to occur (related to off-rate) for a set of related reactions has been studied extensively. Reactions are considered related if the change from one reaction to another is a change in some moieties that does not change the class of reactions (e.g., reactions of amines ($-\text{RNH}_2$) with an acid and varying R groups). For related reactions, the free energy surfaces usually do not intersect. As such, if the equilibrium free energy change is larger for one reaction compared to another, then so is the free energy barrier. Thus, the reaction with the higher affinity will also have a lower off-rate. These relationships are called linear free energy relationships.
in the thymus.

For a specific choice of the interaction energy between the CDR1/CDR2 region of the TCR and the MHC ($E_c$), a panel of one million sequences of TCR peptide contact residues was generated by choosing different amino acids for the variable region according to the frequency with which they appear in the human proteome (results for mouse are in Figure A-6). In the case where there was only one type of self pMHC complex in the thymus, the interaction energy between each TCR and a MHC-bound peptide moiety representative of the human proteome was computed by using the MJ interaction energy matrix and Equation (2.1). Only those TCRs that have interaction energies lying between the positive- and negative-selection thresholds were selected ($|E_p| < |E_{\text{int}}| < |E_n|$). The selected T cells were then challenged with many antigenic peptides characteristic of *L. monocytogenes* (Table 2.1 and (Kosmrlj et al., 2008; Moszer et al., 1995)). A TCR was considered to recognize an antigenic pMHC if the interaction free energy exceeded the recognition threshold, $|E_{\text{int}}| > |E_r|$. Experimental evidence (Naeher et al., 2007) suggests that the negative selection threshold in the thymus is the same as recognition threshold in the periphery, i.e. $E_r = E_n$. In this way, panels of selected T cells that recognize different antigens were generated. Each amino acid on the antigenic peptides was then mutated to the 19 other possibilities, and recognition by the reactive TCRs was again assessed. If more than half the mutations at a particular amino acid site led to abrogation of recognition for an originally reactive T cell, the site was labeled an “important contact”. This procedure was repeated 1,000 times with a different panel of preselection TCRs, and choices for the peptide in the thymus and antigenic peptides to obtain statistics on the number of important contacts characterizing interactions between a typical antigenic peptide and selected TCRs.

For many types of peptides in the thymus, we generated a panel of 10,000 self-peptides using amino acid frequencies characteristic of the human proteome (Table 2.1 and (Flicek et al., 2008; Košmrlj et al., 2008)). The results we obtain are qualitatively robust if at least 100 types of pMHC complexes are in the thymus (Figure A-8). Pathologically large numbers of peptides in the thymus result in deletion
of all thymocytes. Interaction free energies of the panel of TCRs with self-pMHCs were calculated. A TCR was positively selected if it interacted with at least one such pMHC with a free energy that exceeded the positive-selection threshold \( E_p \). To avoid negative selection, a TCR must not interact with any self-pMHC with a free energy that exceeds the negative selection threshold \( E_n \). Important contacts characterizing antigen recognition were determined in the manner described above.

Although the interaction energy between the CDR1 and CDR2 regions of the TCR and MHC \( E_c \) varies continuously as residues on the CDR1 and CDR2 regions change, TCRs can be grouped into three classes based on the relative values of \( E_c \) and the negative-selection threshold, \( E_n \) (Figure 2-2): (i) TCR–MHC interactions are very weak \( (E_c \) and \( E_n \) are separated by a large value); (ii) TCR–MHC interactions are very strong \( (E_c \) and \( E_n \) are separated by a small value); (iii) TCR–MHC interactions are moderate in scale \( (E_c \) and \( E_n \) are separated by a moderate value). Based on recent experimental data (Daniels et al., 2006), for results reported, the difference between \( E_p \) and \( E_n \) is taken to be relatively small \((5k_B T)\).

Very few preselection TCRs with CDR1 and CDR2 loops that interact very weakly with conserved MHC (case i) are positively selected if \( E_p \) and \( E_n \) are relatively close. In effect, they are not MHC restricted. These TCRs are irrelevant for our studies of

---

**Figure 2-2:** Cartoon representation of the three regimes of values of TCR-MHC interactions \( E_c \). In these regimes the TCR–MHC interactions are (i) weak, (ii) strong, and (iii) moderate compared with the threshold for negative selection, \( E_n \). This figure is adapted from Figure 1b in (Košmrlj et al., 2008).
how thymic selection shapes antigen-specific peripheral T cells. TCRs with CDR1 and CDR2 loops that interact very strongly with MHC (\(E_c\) close to or greater than \(E_n\), case ii) are negatively selected with very high probability and so are not relevant for our studies of understanding the origin of how thymic selection results in antigen specificity in the periphery. Not surprisingly then, our studies focus on TCRs with values of \(E_c\) that correspond to moderate interactions between the CDR1/CDR2 loops and MHC (case iii). These TCRs are positively selected with high probability and must avoid negative selection to emerge into the periphery. The particular value\(^3\) of \(E_n - E_c = -40k_BT\) is chosen such that the negative selection deletes \(\sim 80\%\) of positively selected TCRs, when they are selected against many self peptides (van Meerwijk et al., 1997; Merkenschlager et al., 1997). It is enough to study only a single value of TCR binding strength to MHC (\(E_c\)), because a moderate changes in \(E_c\) value do not change the qualitative results (Figure A-7). Therefore having a distribution of TCR binding strengths to MHC also wouldn’t change the qualitative results.

Figure 2-3 shows the frequency of important contacts resulting from our calculations when the conserved TCR–MHC interactions are moderate in scale (\(E_n - E_c\) taken to be \(-40k_BT\) for the results). For TCR selected against many types of peptides, a large fraction of the antigenic peptide’s amino acid sites are important contacts. In contrast, when TCR are selected against one type of peptide in the thymus, very few antigenic peptide amino acid sites are important contacts. This mirrors previous experimental observations (Huseby et al., 2005). We also find that for moderate TCR–MHC interactions, the ability of T cells to mature when only one type of peptide is present in the thymus is limited by positive selection, whereas T cell survival is limited by negative selection when there are many types of peptides in the thymus.

\(^3\)A rough estimate on the bounds of \(E_c\) values of selected TCRs can be obtained without fitting from the condition that the average interaction free energy between TCR and pMHC for selected TCRs is between the thresholds for positive and negative selection \(E_n < E_c + NJ < E_p\), where \(J\) is the average value of interaction between amino acids. The upper (lower) bound \(E_{c,max} = E_p - NJ = -30.5k_BT\) (\(E_{c,min} = E_n - NJ = -35.5k_BT\)) ensures that, on average, interactions result in positive selection and not negative selection. Since for the selection it is enough that a TCR sequence is positively selected by one of many self peptides and avoid being selected by all self peptides the actual bounds for \(E_c\) might be different, but we expect that the range of \(E_c\) values will still be small; viz. \(E_{c,max} - E_{c,min} \propto E_p - E_n\).
Figure 2-3: Selection against many peptides in the thymus results in a larger number of important contacts characterizing antigen recognition. The frequencies of occurrence of one, two, three, etc., important contacts (defined in text) on MHC-bound antigenic peptide moieties recognized by selected TCRs. For TCRs that develop in a thymus with many types of self-peptides (blue curve, $M = 10,000$ peptides) many sites on the antigenic peptide moiety are important contacts. For TCRs that develop in a thymus with only one type of self-pMHC complex (black curve, $M = 1$ peptide) there are far fewer important contacts, indicating less specific (more degenerate or cross-reactive) recognition. This figure is adapted from Figure 1c in (Kosmrlj et al., 2008).

(Figure A-9). Because our computational model recapitulates known experimental data (Figure 2-3 and references (Huseby et al., 2005, 2006)), we used the model to obtain insights into the mechanistic origins of antigen specificity.
2.3 Frustration during negative selection strongly constraints selected TCR sequences.

For a TCR to emerge from the thymus when only one type of pMHC complex is present therein, the binding free energy of the TCR for this pMHC must lie in the interval between $E_n$ and $E_p$. Because the interaction free energy between the TCR's peptide contact residues and the peptide's amino acids is a sum over individual contact free energies (Equation 2.1), many sequences of peptide contact residues on the TCR can satisfy this criterion. A type of selected sequence that occurs with high probability is one where a small number of TCR residues make strong contacts with the corresponding peptide amino acids, and all of the others make irrelevant (i.e., weak) contacts (Figure 2-4). A TCR with such a sequence of peptide contact residues on the TCR would almost certainly be negatively selected when many types of peptides are present in the thymus. This is because it will likely encounter another peptide in the thymus that can differ by only a single amino acid, leading to an additional significant interaction and a total free energy that exceeds $E_n$.

Thus, surviving negative selection presents a frustrating situation because a TCR that avoids negative selection with one peptide in the thymus could be negatively selected by another peptide. Positive selection does not present this problem because, once a TCR receives survival signals by binding a single peptide more strongly than $E_p$, interactions with other peptides are only relevant for negative selection. The frustration associated with subsequently avoiding negative selection by all these diverse pMHCs is the dominant constraint determining peripheral TCR sequences.

To explore how this frustration influences the character of the peptide contact residues of selected TCRs, we developed an analytical approximation (Chapter 3) that suggested that the peptide contact residues on selected TCRs are greatly enriched in amino acids that bind weakly to other amino acids.
Figure 2-4: Schematic description of frustration due to negative selection. The thickness of the bars (or color of peptide amino acids: strong, red; moderate, blue; weak, yellow; very weak, green) is proportional to the interaction energy between TCR and pMHC residues. When developing in a thymus with only one type of endogenous peptide, a TCR that results in a few strong interactions and several weak or moderate interactions with this peptide can survive selection. This is because the total interaction energy falls between the positive- and negative-selection thresholds. The TCR shown in this figure, that survives selection against one type of peptide in the thymus, would likely be negatively selected when there are many types of peptides in the thymus. For example, a peptide that differs by one amino acid from the first one (shown as a change from green to blue colored amino acid) may lead to an additional moderate interaction energy that is sufficient to increase the total interaction energy past the negative selection threshold.

2.4 Negative selection against many self peptides results in TCR sequences with peptide contact residues enriched in weakly interacting amino acids.

To test this suggestion, we first examined the amino acid compositions of the peptide contact residues of the selected TCRs obtained from our computer simulations. When there are many types of peptides in the thymus, peptide contact residues of selected TCRs are enriched in amino acids that interact weakly with other amino acids, whereas strongly interacting amino acids and amino acids with flexible side chains are attenuated (Figure 2-5a). The opposite is true when T cell selection is mediated by a single peptide species in the thymus, with preferential selection of TCR that contain strongly interacting amino acids. In Figure 2-5a, amino acids were ordered according
to the maximum value of the strength with which each amino acid interacts with all others. The nature of the MJ interaction potential is such that this order also reflects the ordering obtained by considering the average value of the interaction free energy of an amino acid with all others. The qualitative results shown in Figure 2-5a are robust to changes in the interaction potential (Figures A-1 - A-4 and Chapter 3). Using different potentials only changes the identities of the amino acids that interact weakly or strongly or the criterion used to define interaction strength. For example, if a potential is such that the order of amino acids obtained by using the average interaction free energies with other amino acids is quite different from that obtained by considering the largest interaction free energies, the qualitative results in Figure 2-5a are obtained if we use the latter quantities to order amino acids.

Do experimental data support our conclusion that frustration due to negative selection skews the mature T cell repertoire to TCRs composed of peptide contact residues enriched in amino acids that bind weakly to other amino acids? We analyzed the 18 available crystal structures of TCR bound to class I pMHC complexes to obtain the frequency with which different amino acids are represented at residues of the TCR that contact the peptide (Kaas et al., 2004). All TCR moieties that contact peptide amino acids were considered, and two methods were used to identify these contact residues. One was to define a contact as a position where a water molecule does not fit in the gap between a TCR residue and a peptide amino acid. In the other method, residues in contact have their \( C_\alpha \) atoms within 6.5 Å of each other. The qualitative results are the same for both methods (Figure A-10), and in Figure 2-5b, we show results using the second criterion.

Whereas the qualitative computational results (Figure 2-5a) are independent of interaction potential, to compare the experimental data with this prediction, we need to know whether a particular amino acid is weak or strong in reality. We have used two different prescriptions to order the amino acids according to the strength of their interactions with other amino acids. One is to use the MJ matrix, but the order thus obtained has been criticized because it overemphasizes hydrophobic interactions and considers interactions between charged amino acids to be weak (Sorenson and
Figure 2-5: Selection against many types of peptides in the thymus results in selected TCRs with peptide contact residues with an enhanced frequency of amino acids that interact weakly with all other amino acids. The ordinate is the ratio of the frequencies of occurrence of an amino acid in the peptide contact residues of selected TCRs to preselection TCRs. (a) For the computational results, the abscissa is a list of amino acids ordered according to the maximum free energy (as per the MJ interaction potential) with which it interacts with all other amino acids. The qualitative results are robust to changes in potential (Figures A-1 – A-4). (b) The ordinate was obtained by analyzing the 18 available crystal structures of TCR-pMHC (class I) complexes as described in the text. Amino acids were classified as strongly interacting (IVYWREL) or weakly interacting (QSNTAG) following the reference (Zeldovich et al., 2007). This figure is adapted from Figure 2 in (Kosmrlj et al., 2008).

Head-Gordon, 1998). Data obtained by examining the stability of thermophiles are proposed to be better suited for analyzing the strength of interactions between amino acids (Zeldovich et al., 2007), and posit that the strongly interacting amino acids are IVYWREL, and the weakly interacting ones are QNSTAG (Zeldovich et al., 2007).

Figure 2-5b shows results where amino acids are divided into two classes (weak and strong) according to this prescription. The data obtained from crystal structures are in qualitative agreement with the theoretical prediction in that weakly interacting amino acids are enriched on peptide contact residues of the TCR, and strongly interacting amino acids are attenuated. Using the MJ matrix leads to similar results.
(Figure A-10), except that charged amino acids (R, E, K), which are weak according
to the MJ matrix, are additional outliers. Tyrosine is considered to be a strongly in-
teracting amino acid by either approach, but is well represented in the TCR–peptide
contact residues. This may be because a germ-line-encoded tyrosine interacts with
a conserved MHC residue that is close to the peptide amino acids (Dai et al., 2008;
Feng et al., 2007), and so it may interact ubiquitously with peptide amino acids.

Our results suggest that negative selection against many types of thymic peptides
results in mature TCRs with peptide contact residues that interact weakly with other
amino acids. How does this influence their antigen specificity?

2.5 Antigen specificity is the result of TCR residues
binding peptides via multiple moderate inter-
actions.

In our model, the interaction energy between an antigenic peptide and residues of
a TCR that recognizes it is the sum of 10 numbers, with each number being the
interaction energy between an amino acid on the peptide and the corresponding TCR
contact site (Figure 2-1). We computed the values of these site–site interaction en-
ergies using all our TCR–antigenic peptide pairs. In Figure 2-6, we compare the
frequency with which each value of these interaction energies occurs for three cases:
preselection TCRs, TCRs that developed in a thymus with many types of pMHC, and
TCRs that developed in a thymus with one type of pMHC.

Our results indicate that, compared with the preselection TCRs, antigen recog-
nition by TCRs selected against many types of pMHC complexes is mediated by
fewer strong and weak amino acid–amino acid interactions, resulting in a pronounced
enhancement of moderate interactions. This result is consistent with experimental
observations of Savage and Davis (2001). This focusing on moderate interactions
is because negative selection constrains mature TCR peptide contact residues to be
composed of weakly interacting amino acids (Figure 2-5). Therefore the selected
Figure 2-6: Distribution of amino acid–amino acid contact energies (in units of $k_B T$, described in text) characterizing interactions between selected reactive TCRs and antigenic peptides suggest the basis for specificity. The distribution of interaction energies between individual amino acids on peptide contact residues on the TCR and antigenic peptides are shown. The distribution for TCRs that develop in a thymus with many endogenous peptides (blue curve) is very different from that for preselection TCRs (red curve). The distribution of contact energies is not significantly altered for TCRs that develop in a thymus with only one type of peptide (black curve) compared with preselection TCRs. This figure is adapted from Figure 3 in (Kosmrlj et al., 2008).

TCRs can only recognize, i.e. bind sufficiently strongly to, antigenic peptides that are composed of predominantly strongly interacting amino acids and amino acids with flexible side chains (Figure 2-7a). To test this experimentally, we analyzed amino acid frequencies in a list of best known HIV peptides (Llano et al., 2009) and compared them to amino acid frequencies in HIV proteome (Belshaw et al., 2009). Figure 2-7b shows that HIV peptides are indeed enriched with strongly interacting amino acids. Thus the weakly interacting amino acids on the TCR bind to strongly interacting amino acids on antigenic peptides resulting in multiple moderate scale interactions.
Figure 2-7: Amino acid frequencies of recognized antigenic peptides are enriched with strongly interacting amino acids. (a) For the computational results, the ordinate shows is the ratio of amino acid frequencies of antigenic peptides that are recognized by at least one of the selected TCRs with respect to amino acid frequencies of all antigenic peptides. The black curve depicts the results for TCRs selected against one self-peptide \( (M = 1) \), whereas the blue curve corresponds to selection against many self peptides \( (M = 10,000) \). The amino acids on the abscissa are ordered from strongest (L) to weakest (K) according to the strongest interaction with another amino acid in the MJ matrix. (b) The ordinate was obtained by analyzing a list of the best known recognized HIV peptides (Llano et al., 2009) and the HIV proteome (Belshaw et al., 2009). Amino acids were classified as strongly interacting (IVYWREL) or weakly interacting (QSNTAG) following the reference (Zeldovich et al., 2007).

that add up to a total binding free energy that is large enough for recognition. Because antigen recognition is mediated by multiple interactions of moderate value, each contact makes a significant contribution to the total interaction free energy necessary for recognition. Therefore, disrupting any interaction by mutating one of the strongly interacting amino acids on the peptide results in abrogation of recognition (Figure 2-8). This is the origin of antigen specificity. This prediction is consistent with measurements reported for the B3K 506 TCR, which was selected against many types of pMHC complexes in the thymus and recognizes the 3K–IA\(^b\) pMHC (Huseby et al., 2006). Many mutations of the antigenic peptide correspond to moderate \( \Delta \Delta G \)
Figure 2-8: Mechanism for specificity and degeneracy of TCR recognition of antigenic peptides. The weakly interacting amino acids (green) on the TCR bind to strongly interacting amino acids (red, blue) on antigenic peptides resulting in multiple moderate scale interactions that add up to a total binding free energy that is large enough for recognition. Because antigen recognition is mediated by multiple interactions of moderate value, each contact makes a significant contribution to the total interaction free energy necessary for recognition. Therefore, disrupting any interaction by mutating one of the strongly interacting amino acids on the peptide results (shown as a change from red to yellow color) in abrogation of recognition. At the same time TCR recognition of antigenic peptides is degenerate, because there are many combinatorial ways of distributing strongly interacting amino acids (red, blue) along the peptide, which results in a sufficiently strong binding with TCR for recognition.

values, and each contributes significantly to recognition. At the same time TCR recognition of antigenic peptides is degenerate, because there are many combinatorial ways of distributing strongly interacting amino acids along the peptide, which results in a sufficiently strong binding with TCR for recognition (Figure 2-8). TCRs that survive negative selection against many types of peptides are quite diverse because many sequences are consistent with the constraint that peptide contact residues are predominantly composed of amino acids that interact weakly with all others.

When there is one type of pMHC complex in the thymus, the peptide-binding residues of selected TCRs are not subject to the important constraints of avoiding negative selection against many types of peptides, and moderate amino acid–amino acid interactions do not dominate (Figure 2-6). Figure 2-5a suggests that strongly interacting amino acids are represented more than in the preselection repertoire. However, the value of TCR binding strength to MHC, $E_c$, was tuned for the selection
against many self peptides. If we allow the distribution of $E_c$ values, the TCRs that bind more strongly to MHC (larger $|E_c|$ values) may also survive the thymic selection against only one type of self peptide (Figure A-9) and such TCRs have a lower frequency of strongly interacting peptide contact residues (see Chapter 3 and Figure A-7). Therefore, sufficiently strong binding of TCR for antigenic peptide recognition may be reached by a few strong interactions and by stronger binding to MHC. Thus, mutating the antigenic peptide amino acids that contact strongly interacting amino acid residues on the TCR should abrogate recognition, but mutations at most other sites should have little impact. This is reflected in the experimental data reported by Huseby et al. (2006). For one example, consider the YAe62.8 TCR, which is selected against a single type of peptide in the thymus and recognizes variants of the 3K–IA$^b$ antigenic peptide. Most mutations to the antigenic peptide result in small changes in $\Delta \Delta G$, but one mutation results in a large change. This one major peptide contact dominates the interaction free energy with the others being irrelevant, and this is the origin of enhanced cross-reactivity. It may also happen that the binding free energy between a TCR and antigenic peptide is sufficiently strong that a single mutation of peptide amino acids cannot prevent recognition, which results in 0 important contacts (Figure 2-3).

2.6 Escape of T cells, which could trigger an autoimmune disease

Thymic selection is not perfect and it may happen that some TCRs, which interact strongly with self pMHCs, escape from the thymus. Such TCRs could potentially lead to development of autoimmune disease, and in the remainder of this Chapter they are labeled as “autoimmune” TCRs. Due to stochastic effects, it may happen that a diffusing T cell in the thymus never interacts with some peptides that would lead to negative selection (see Chapter 5). It may also happen that the negative selection is not perfect, i.e. even if a TCR binds strongly to self-pMHC it can escape with some
probability. Here we only focus on the latter effect, which can be modeled with a soft threshold for negative selection. For a TCR $\tilde{t}$ that interacts with self peptide $\tilde{s}$, the probability of negative selection is assumed to be

$$P_n(\tilde{t}, \tilde{s}) = \frac{1}{1 + \exp \left( -\frac{(E_{\text{int}}(\tilde{t}, \tilde{s}) - E_n)}{\sigma_n} \right)},$$  \hspace{1cm} (2.2)

where the parameter $\sigma_n$ denotes the softness of negative selection threshold. For a TCR that interacts strongly with self-pMHC ($|E_{\text{int}}| > |E_n|$), the probability of negative selection is close to 1, while for a TCR that interacts weakly with self-pMHC ($|E_{\text{int}}| < |E_n|$) the probability of negative selection is small. Similarly we can define the probability for positive selection $P_p$ with the corresponding softness $\sigma_p$. From experiments (Daniels et al., 2006) we know that the threshold for positive selection is softer ($\sigma_p > \sigma_n$).

In this casethymic selection is modeled by testing each TCR sequence with all self-pMHCs: for each self-peptide we calculate the corresponding probabilities $P_p$ and $P_n$ of positive and negative selection, then we draw two uniformly distributed random numbers $r_p$ and $r_n$ on the $(0, 1)$ interval and a TCR is positively (negatively) selected if $r_p < P_p$ ($r_n < P_n$). After thymic selection is completed, we check if any of the selected TCRs interacts strongly with any self-peptide ($|E_{\text{int}}| > |E_n|$). Deterministic criteria are now used because strong interaction free energy leads to high probability of T cell activation. Any other deterministic or stochastic criteria would not qualitatively change the results. We find that the introduction of soft thresholds for positive and negative selection does not qualitatively change the results reported in the first part of this chapter (Figure A-5).

Figure 2-9a shows that increasing the softness of the threshold for negative selection $\sigma_n$, increases the chance of escape of autoimmune TCRs. This is because strongly interacting TCRs are negatively selected with lower probability when threshold for negative selection is softer. Interestingly the ratio of the numbers of autoimmune T cells to the number of selected T cells seem to be roughly constant with the number $M$ of self-peptides used during the development in thymus (Figure 2-9b). The fraction of
Figure 2-9: Ratio of the numbers of escaped autoimmune TCRs and selected TCRs as a function of: (a) the softness of threshold for negative selection $\sigma_n$, (b) number of self peptides $M$. The fraction of autoimmune TCRs increases with $\sigma_n$, until the softness of the thresholds becomes of the same order as the separation between the thresholds of positive negative selection, $\sigma_n + \sigma_p \sim |E_n - E_p|$. The errorbars in (a) and (b) correspond to the standard deviation of the fractions of escaped TCRs obtained from repeating the thymic selection process many times. $(E_n - E_c = -21k_BT, E_n - E_p = -2.5k_BT, N = 5, M = 1,000, \sigma_n = 0.1k_BT, \sigma_p = 1k_BT)$.

Autoimmune T cells first increases for small $M$, but then it becomes roughly constant. This may be because the T cell repertoire properties in the EVD calculation scale as $\sqrt{\ln M}$ (see Chapter 3). Note that with increasing number of self-peptides $M$, both the number of autoimmune T cells and the number of selected T cells decreases, but the ratio is roughly constant. While this ratio is roughly independent of $M$ in the relevant regime ($M \sim$ a few thousands, Figure 2-9b), the absolute numbers of escaped autoimmune T cells is higher for smaller $M$. This implies that the rate for escape of autoimmune T cells could be higher in people expressing MHCs that bind fewer types of self peptides. This may be why people expressing HLA-B57 that binds $\sim$2-3 times fewer self peptides, and are better at controlling HIV infection (see Chapter 4), are also more prone to autoimmunity.

Experimental results show that the escape of autoimmune TCRs is higher in
humans compared to mice, and the results of Figure 2-9a suggest that this may be
due to the differences in the softness of threshold for negative selection.

2.7 Conclusions

Although important clues were provided by the experimental data reported by Huseby
et al. (2005, 2006), a mechanistic understanding of how thymic selection designs TCR
sequences that are simultaneously antigen specific, cross-reactive, diverse, and self-
tolerant remained unclear. Our computational studies shed light on these issues.

If a TCR receives survival signals from a self-pMHC complex, it is positively
selected. Interactions with the other peptides expressed in the thymus are then only
relevant for negative selection. Positive selection ensures MHC restriction, enables
weak binding of TCRs to self pMHC, and influences the fraction of T cells that
survive thymic selection. Thus, it mediates important properties. However, antigen
specificity appears to be determined by the requirement that positively selected T
cells must survive negative selection.

TCR sequences must simultaneously avoid being negatively selected by many
endogenous MHC-bound peptides, and this imposes strong constraints on the nature
of the peptide contact residues of selected TCRs. We find that this is why, in mature T
cells, these residues are enriched in amino acids that interact weakly with other amino
acids (referred to as weak amino acids). For a selected TCR to recognize an antigenic
peptide in the periphery, it must bind to it with an affinity that exceeds a threshold.
This can occur only if the peptide is composed of amino acids that are among the
strongest binders of the corresponding weak amino acids of the TCR’s peptide contact
residues (Figure 2-7), resulting in a number of moderate scale interactions that sum
up to exceed the threshold affinity required for recognition. Because each moderate
interaction contributes a substantial fraction of the overall affinity, disrupting most
of them (via mutations) abrogates recognition. Thus, antigen specificity emerges
because TCR residues that contact the peptide are enriched in amino acids that
interact weakly with other amino acids. It is worth remarking that weakly binding
amino acids are not always the mediators of recognition; TCR selected against one type of peptide do not exhibit this behavior (Huseby et al., 2005, 2006).

Because the amino acids treated explicitly in our model include variable MHC residues, our results are also consistent with data showing that TCR selected against many peptides are also MHC specific. We note in passing that we have also studied the alloreactivity of selected TCRs, i.e. reactions against nonself MHC molecules (data not shown). Our findings suggest that the relative importance of the peptide (compared with the MHC) in mediating alloreactive responses depends on how different the allo- and endogenous MHCs are vis-à-vis their interaction energies with the CDR1 and CDR2 loops of a particular TCR ($E_c$ in our model); the greater this difference, the less important the peptide.

Our results suggest a model for specificity of TCR–antigenic pMHC recognition that is different from Fisher’s lock and key metaphor for the specificity with which an enzyme binds its substrate (Fischer, 1894). It also appears to be different from that applicable to specificity of antibody–antigen interactions where shape complementarity and multiple weak interactions are inextricably coupled (Perelson and Oster, 1979). Shape complementarity is important for TCR recognition of antigen in two ways (Figure 2-10). First, it plays a key role in peptide binding to the MHC groove, and hence influences antigen presentation. Secondly, shape complementarity is possibly important in mediating interactions of the TCR with MHC moieties, which results in orienting the TCR in a way that juxtaposes its peptide contact residues with the peptide. Indeed, it has been suggested that if the peptide has a conformation that is not relatively flat, it disrupts TCR–MHC interactions, thereby preventing positive selection (Schumacher and Ploegh, 1994). But, these TCR–MHC interactions required for positive selection and binding of peripheral TCRs to MHC in the proper orientation do not confer peptide specificity.

Once properly oriented, a TCR scans the relatively flat conformation of the short peptide, and recognizes the epitope if a number of peptide amino acids correspond to strong binders for the weak peptide contact residues of this TCR. For reasons described above, recognition is specific because each resulting interaction is moderate.
Figure 2-10: A bar code scanning model for specificity of TCR recognition of antigenic peptides. The thickness of the lines in the cartoon is proportional to the strength of TCR–peptide interactions. This figure is adapted from Figure 4 in (Košmrlj et al., 2008).

Shape complementarity seems to be decoupled from the origin of specificity. TCR recognition of antigen is analogous to scanning a flat bar code for the appropriate number of moderately thick lines. In this metaphor, the moderately thick lines represent moderate interactions mediated by peptide amino acids that are strong binders for the weak amino acids that comprise the TCR’s peptide contact residues. This bar-code model also makes vivid why specificity and cross-reactivity can coexist. For example, consider a situation where any three of four contacts with the peptide amino acids need to be of moderate scale for recognition; i.e., three of the four lines need to be moderately thick. If a particular peptide satisfies this criterion (say, lines 1, 3, and 4 are moderately thick), mutations at any one of these sites will abrogate recognition (specificity). But another peptide that leads to lines 1, 2, and 3 being moderately thick will also be recognized by this TCR (cross-reactivity). One might say that TCRs scan a bar code and recognize statistical patterns — ones that have a sufficient number of moderately thick lines.

We hope that the results we have reported will motivate experimental and computational studies that will ultimately elucidate how one of nature’s intriguing designers (the thymus) works and how its aberrant regulation can contribute to autoimmune
disease. An important question unresolved by our studies is how variability in expression levels of different types of endogenous peptides in the thymus influences the T cell repertoire.
Chapter 3

Thymic selection of T-cell receptors as an extreme value problem

The adaptive immune system clears pathogens from infected hosts with the aid of T lymphocytes (T cells). Foreign (antigenic) and self-proteins are processed into short peptides (p) inside antigen-presenting cells (APC), bound to MHC proteins, and presented on the surface of APCs. Each T-cell receptor (TCR) has a conserved region participating in the signaling functions, and a highly variable segment responsible for antigen recognition. Because variable regions are generated by stochastic rearrangement of the relevant genes, most T cells express a distinct TCR. The diversity of the T cell repertoire enables the immune system to recognize many different antigenic short pMHC complexes. Peptides presented on MHC class I are typically 8–11 amino acids long (Kuby et al., 2006), which is enough to cover all possible self-peptides (the human proteome consists of \( P \approx 10^7 \) amino-acids (Burroughs et al., 2004; Flicek et al., 2008)) as well as many antigenic peptides. TCR recognition of pMHC is both specific and degenerate (Figure 1-2). It is specific, because most mutations to the recognized peptide amino acids abrogate recognition (Huseby et al., 2006, 2005). It is degenerate because a given TCR can recognize several antigenic peptides (Hemmer

Part of this chapter is based on the publication (Košmrlj et al., 2009).
The gene rearrangement process ensuring the diversity of TCR is random. It may thus result in T cells potentially harmful to the host, because they bind strongly to self peptide-MHC complexes; or useless T cells which bind too weakly to MHC to recognize antigenic peptides. Such aberrant TCRs are eliminated in the thymus (von Boehmer et al., 2003; Werlen et al., 2003; Siggs et al., 2006; Hogquist et al., 2005), where immature T cells (thymocytes) are exposed to a large set ($10^3 - 10^4$) of self-pMHC (see also subsection 1.1.3 and Figure 1-3). Thymocytes expressing a TCR that binds with high affinity to any self-pMHC molecule are deleted in the thymus (a process called negative selection). However, a thymocyte’s TCR must also bind sufficiently strongly to at least one self pMHC complex to receive survival signals and emerge from the thymus (a process called positive selection).

### 3.1 Thymic selection model

Signaling events, gene transcription programs, and cell migration during T cell development in the thymus (von Boehmer et al., 2003; Werlen et al., 2003; Siggs et al., 2006; Hogquist et al., 2005; Daniels et al., 2006; Bousso et al., 2002; Borghans et al., 2003; Scherer et al., 2004; Detours et al., 1999; Detours and Perelson, 1999) have been studied extensively. Despite many important advances, how interactions with self-pMHC complexes in the thymus shape the peptide-binding properties of selected TCR amino acid sequences, such that mature T cells exhibit their special properties, is poorly understood. To address this issue, in Chapter 2 and reference (Košmrlj et al., 2008) we numerically studied a simple model where TCRs and pMHC were represented by strings of amino acids (Figure 2-1). These strings indicate the amino-acids on the interface between TCRs and pMHC complexes, and it is assumed that each site on a TCR interacts only with a corresponding site on pMHC. The binding interface of TCR is actually composed of a region that is in contact with the MHC molecule, and a segment that is in contact with the peptide. It is the latter part that
is highly variable, while the former is more conserved. We shall therefore explicitly consider only the former amino-acids, but not the latter. Similarly, there are many possible peptides that can bind to MHC, and their sequences are considered explicitly, whereas those of the MHC are not. We could in principal add a few sites to the TCR and pMHC strings to account for any variability in the segments not considered.

Simplified representations of amino-acids (e.g., as a string of numbers or bits) were employed earlier (Detours et al., 1999; Detours and Perelson, 1999; Chao et al., 2005) in the context of TCR-pMHC interactions, mainly to report that negative selection reduces TCR cross-reactivity. In Chapter 2 and reference (Kosmrlj et al., 2008), we numerically studied the model in Figure 2-1 (and described below) to qualitatively describe the role of positive and negative selection on the amino-acid composition of selected TCRs. By randomly generating TCR and pMHC sequences, and implementing thymic selection in silico, we showed that selected TCRs are enriched in weakly interacting amino acids, and explained how this leads to specific, yet cross-reactive, TCR recognition of antigen, a long-standing puzzle (see Chapter 2 and reference (Kosmrlj et al., 2008)). In this chapter we show that the model can be solved exactly in the limit of long TCR/peptide sequences. The resulting analytic expression for the amino-acid composition of selected TCRs is surprisingly accurate even for short peptides and provides a theoretical basis for previous numerical results. Furthermore, we are able to obtain a phase diagram that indicates the ranges of parameters where negative or positive selection are dominant, leading to quite different bias in selection/function.

To assess the effects of thymic selection, as well as antigen recognition, we evaluate the free energy of interaction between TCR-pMHC pairs (for brevity, free energy will be referred to as energy). The interaction energy is composed of two parts: a TCR interaction with MHC, and a TCR interaction with the peptide. The former is given a value $E_c$ (which may be varied to describe different TCRs and MHCs). The latter is obtained by aligning the TCR and pMHC amino-acids that are treated explicitly, and adding the pairwise interactions between corresponding pairs. For a given TCR-
pMHC pair, this gives
\[ E_{\text{int}}(\vec{t}, \vec{s}) = E_c + \sum_{i=1}^{N} J(t_i, s_i), \tag{3.1} \]

where \( J(t_i, s_i) \) is the contribution from the \( i \)th amino acids of the TCR \((t_i)\) and the peptide \((s_i)\), and \( N \) is the length of the variable TCR/peptide region. The matrix \( J \) encodes the interaction energies between specific pairs of amino-acids. For numerical implementations we use the Miyazawa-Jernigan (MJ) matrix (Miyazawa and Jernigan, 1996) that was developed in the context of protein folding.

Immature T cells interact with a set \( S \) of \( M \) self-pMHC complexes, where typically \( M \) is of the order of \( 10^3 - 10^4 \). To mimic thymic selection, sequences that bind to any self-pMHC too strongly \((E_{\text{int}} < E_n)\) are deleted (negative selection). However, a thymocyte’s TCR must also bind sufficiently strongly \((E_{\text{int}} < E_p)\) to at least one self-pMHC to receive survival signals and emerge from the thymus (positive selection). A thymocyte expressing TCR with string \( \vec{t} \) will thus be selected if the strongest interaction with self-pMHC is between thresholds for negative and positive selection, i.e.

\[ E_n < \min_{\vec{s} \in S} \{ E_{\text{int}}(\vec{t}, \vec{s}) \} < E_p. \tag{3.2} \]

Recent experiments (Daniels et al., 2006) show that the difference between thresholds for positive and negative selection is relatively small (a few \( k_BT \)).

### 3.1.1 Extreme value problem

Equation (3.2) casts thymic selection as an extreme value problem (Leadbetter et al., 1983), enabling us to calculate the probability \( P_{\text{sel}}(\vec{t}) \) that a TCR sequence \( \vec{t} \) will be selected in the thymus. Let us indicate by \( \rho(x|\vec{t}) \) the probability density function (PDF) of the interaction energy between the TCR \( \vec{t} \) and a random peptide. The PDF \( \Pi(x|\vec{t}) \) of the strongest (minimum) of the \( M \) independent random interaction energies is then obtained by multiplying \( \rho \) with the probability of all remaining \((M - 1)\) energy values being larger – \( (1 - P(E < x|\vec{t}))^{M-1} \), where \( P(E < x|\vec{t}) \) is the cumulative probability – and noting the multiplicity \( M \) for which energy is the lowest. The
probability that TCR \( \vec{t} \) is selected is then obtained by integrating \( \Pi(x|\vec{t}) \) over the allowed range, as

\[
P_{\text{sel}}(\vec{t}) = \int_{E_n}^{E_p} \Pi(x|\vec{t}) \, dx, \quad \text{with}
\]
\[
\Pi(x|\vec{t}) = M \rho(x|\vec{t}) \left(1 - P(E < x|\vec{t})\right)^{M-1}.
\] (3.3)

For \( M > 1 \), this extreme value distribution (EVD) converges to one of three possible forms, (Leadbetter et al., 1983) depending on the tail of the PDF for each entry. Equation (3.1) indicates that in our case as each energy is the sum of \( N \) contributions, \( \rho(x|\vec{t}) \) should be a Gaussian for large \( N \), in which case the relevant EVD is the Gumbel distribution. (Leadbetter et al., 1983)

To obtain an explicit form for \( \Pi(x|\vec{t}) \), we model the set \( S \) of self-peptides as \( M \) strings in which each amino-acid is chosen independently. The probability \( f_a \) for selecting amino-acid \( a \) at each site is taken to be the frequency of this amino-acid in the self-proteome. For a specific TCR sequence \( \vec{t} \), the average interaction energy with self peptides follows from Equation (3.1) as

\[
E_{\text{av}}(\vec{t}) = E_c + \sum_{i=1}^{N} \mathcal{E}(t_i),
\]
where we have denoted the average over self amino-acid frequencies by \( [G(a)]_a \equiv \sum_{a=1}^{20} f_a G(a) \). Similarly, the variance of the interaction energy is

\[
\mathcal{V}(\vec{t}) = \sum_{i=1}^{N} \mathcal{V}(t_i),
\]
where \( \mathcal{V}(t_i) = [J(t_i, a)]_a^2 - [J(t_i, a)]_a \). For large \( N \), we can approximate \( \rho(x|\vec{t}) \) with a Gaussian PDF with the above mean and variance. From standard results for the Gumbel distribution (Leadbetter et al., 1983), we conclude that in the limit of \( M \gg 1 \), the peak of the distribution \( \Pi(x|\vec{t}) \) is located at

\[
E_0(\vec{t}) = E_{\text{av}}(\vec{t}) - \sqrt{2\mathcal{V}(\vec{t}) \ln M},
\] (3.4)

and its width is \( \Sigma_0(\vec{t}) = \sqrt{\pi^2 \mathcal{V}(\vec{t})/(12 \ln M)} \). (Since the PDF \( \rho(x|\vec{t}) \) originates from a bounded set of energies, it is strictly not Gaussian in the tails. Hence, once the extreme values begin to probe the tail of the distribution, the above results will no longer be valid. Indeed, in the limit when \( M \sim \mathcal{O}(20^N) \), the EVD will approach a delta-function centered at the \( M \)-independent value corresponding to the optimal
binding energy.)

3.1.2 Hamiltonian minimization

In the limit of long TCR/peptides ($N \gg 1$), we can exactly calculate the statistics of the amino-acid composition of selected TCRs. To obtain a proper thermodynamic limit, we need to set $\{E_c, E_p, E_n\} \propto N$, and $\ln M \propto N$. The latter ensures that the peak of the distribution, $E_0(\tilde{t})$, is proportional to $N$, and also results in a width $\Sigma_0(\tilde{t})$ which is independent of $N$. (The relation $\ln M = \alpha N$ can be justified with the expectation that $M$ should grow proportionately to the proteome size $P$, while $N \propto \ln P$ to enable encoding the proteome.) In this large $N$ limit, the EVD is sufficiently narrow that the value of the optimal energy can be precisely equated with the peak $E_0(\tilde{t})$, and Equation (3.2) for the selection condition can be replaced with

$$E_n < E_c + \sum_{i=1}^{N} E(t_i) - \sqrt{2 \ln M \sum_{i=1}^{N} V(t_i) < E_p}.$$

(3.5)

Thus, for each sequence $\tilde{t}$, we have to evaluate the ‘Hamiltonian’ $E_0(\tilde{t})$, and the sequence is accepted if this energy falls in the interval $(E_n, E_p)$. This is somewhat similar to the micro-canonical ensemble in Statistical Physics, with the restriction of the energy to an interval rather than a narrow range only a minor elaboration (see below). From the equivalence of canonical and micro-canonical ensembles for large $N$, we know that the probability for a sequence is governed by the Boltzmann weight $p(\tilde{t}) \propto \left(\prod_{i=1}^{N} f_{t_i}\right) \exp[-\beta E_0(\tilde{t})]$. Here $\{f_{t_n}\}$ indicate the natural frequencies of the different amino-acids prior to selection, while the effect of thymic selection is captured in the parameter $\beta$ which is determined by solving for the average energy.

The presence of $\sqrt{2 \ln M \sum_{i} V(t_i)}$ in the Hamiltonian initially appears as a complication that makes exact computation of the average energy from $\exp[-\beta E_0(\tilde{t})]$ impossible. However, this apparent ‘coupling’ is easily dealt with by standard methods such as Legendre transforms or Hamiltonian minimization (Kardar, 1983). This can be justified easily as follows: We need to solve a ‘Hamiltonian’ $\mathcal{H}(U, V)$ which
Figure 3-1: Schematic presentation of the choice of parameter $\beta$ that constrains the average energy $\langle E_0 \rangle$ to the range in Equation (3.5), while maximizing entropy. Three cases above correspond to the examples where (a) $\beta > 0$, (b) $\beta = 0$ and (c) $\beta < 0$.

depends on two extensive quantities $U = \sum_{i=1}^{N} E(t_i)$ and $V = \sum_{i=1}^{N} V(t_i)$. The corresponding partition function can be decomposed as $Z = \sum_{V} \Omega(U, V) e^{-\beta H(U, V)}$, but can be approximated with its largest term. Note that the same density of states $\Omega(U, V) \equiv e^{S(U, V)/k_B}$ appears, irrespective of the specific form of $H(U, V)$. In particular, the choice $H_0 = E_c + U - \gamma V - \ln M/(2\gamma) = E_c + \sum_{i=1}^{N} (E(t_i) - \gamma V(t_i)) - \ln M/(2\gamma)$ corresponds to a set of non-interacting variables, with

$$p(\tilde{t}) \propto \prod_{i=1}^{N} f(t_i, \exp[-\beta(E(t_i) - \gamma V(t_i))]),$$

(3.6)

for which thermodynamic quantities (such as entropy) are easily computed. By judicious choice of $\gamma$ we can then ensure that the same average energy appears for $H_0(\tilde{t})$ and our $E_0(\tilde{t})$. Using Legendre transforms, which is equivalent to minimizing $H_0(\tilde{t})$ with respect to $\gamma$, one finds that the required $E_0(\tilde{t})$ is obtained by setting $\gamma(\beta) = \sqrt{\ln M/(2N \langle V \rangle_{\beta, \gamma})}$, where $\langle \cdots \rangle_{\beta, \gamma}$ refers to the average with the non-interacting weight $e^{-\beta(U - \gamma V)}$.

Finally, the value of $\beta$ has to be determined by constraining the average energy determined above to the range in Equation (3.5), while maximizing entropy (Figure 3-1). Given the bounded set of energies, the inverse temperature $\beta$ can be either negative or positive. The $2^N$ possible values for $E_0(\tilde{t})$ span a range from $E_{\min}$ to $E_{\max}$, and a corresponding number of states $\Omega(E_0)$ which is a bell-shape between
Figure 3-2: Color representation of the dependence of the inverse temperature $\beta$ on the number of self-peptides $\ln M/N$ and the threshold for negative selection energy $E_n/N$ with $(E_p - E_n)/N = 0.5k_BT$ in the limit of large $N$. The region between the black lines corresponds to $\beta = 0$, to the right (left) of which negative (positive) selection is dominant, and weak (strong) amino-acids are selected. Note that as $(E_p - E_n)/N$ goes to zero, the intermediate region disappears. The dotted lines indicate the relevant parameter values for thymic selection in mouse (see text) that result in $\beta = -0.37(k_BT)^{-1}$. This figure is adapted from Figure 2 in (Košmrlj et al., 2009).

these extremes with a maximum at some $E_{\text{mid}}$. If $E_{\text{mid}} > E_p$, we must set $\beta$ such that $\langle E_0(\vec{t}) \rangle = E_p$. In this case, $\beta > 0$, positive selection is dominant and stronger amino-acids are selected. If $E_{\text{mid}} < E_n$, we must set $\beta$ such that $\langle E_0(\vec{t}) \rangle = E_n$, $\beta < 0$, negative selection is dominant and weaker amino-acids are selected. For $E_n < E_{\text{mid}} < E_p$, we must set $\beta = 0$ and there is no modification due to selection.

Figure 3-2 depicts the variation of $\beta$ as a function of $\ln(M)/N$ and threshold for negative selection $E_n$ with $(E_p - E_n)/N = 0.5k_BT$. For the set of parameters that are relevant for thymic selection in mouse (Chapter 2 and (Košmrlj et al., 2008)), i.e. $N = 5$, $E_n - E_c = -21k_BT$, $E_p - E_n = 2.5k_BT$ and $M = 10^3$, we find $\beta = -0.37(k_BT)^{-1}$ which means that negative selection is dominant and weaker amino-
acids are selected. Also $\gamma = 0.83(k_B T)^{-1}$ indicating a preference for amino-acids with smaller variations in binding energy. With these parameters we can calculate the amino-acid frequencies of selected TCRs as

$$f^{(\text{sel})}_{a} = \frac{f_a \exp \left[ - \beta (E(a) - \gamma V(a)) \right]}{\sum_{b=1}^{20} f_b \exp \left[ - \beta (E(b) - \gamma V(b)) \right]}.$$  \hspace{1cm} (3.7)

It is important to ask if the above expression, exact in the limit of $N \to \infty$, has any relevance to the actual TCR/peptides with $N \sim 5 - 10$. We thus numerically simulated the case of $N = 5$ by generating a random set of $10^6$ TCR sequences, and selected them against $M = 10^3$ self-peptides. The selected TCRs were used to construct the amino-acid frequencies depicted in Figure 3-3. The red line in this figure comes from Equation (3.7) with the same $J$ and $\{f_a\}$. The agreement between the two is remarkable given the small value of $N = 5$, and may be indicative of small corrections to the $N \to \infty$ result. Since $\beta < 0$ for the relevant parameters, the positive selection has no effect in the limit of $N \to \infty$. To test the effect of positive selection in simulations, we have repeated the simulations of thymic selection with only the negative selection process (i.e. every TCR is positively selected or $(E_p - E_n)/N \to -\infty$). Amino acid frequencies of thus selected TCRs are presented with green data points on Figure 3-3 and they are similar to TCRs (black data points), which were selected with both the negative and positive selection processes. The one effect we notice, is that the removal of positive selection (green data points) makes the errorbars smaller and thus the analytical result, becomes less accurate ($\sim 50\%$) for strongly interacting amino acids (e.g. L). The reason for large discrepancy is likely due to the differences between the micro-canonical and canonical ensembles, which should be more pronounced for small values of $N$.

### 3.2 T cell recognition of foreign peptides

Once T cells complete the thymic selection, a set $\mathcal{T}$ of TCRs ($K$ in number) is released in the blood stream, where they try to identify infected cells. A T cell
recognizes infected cells when its TCR binds sufficiently strongly ($E_{\text{int}} < E_r$) to foreign peptide-MHC. Experimental evidence (Naeher et al., 2007) suggests that the negative selection threshold in the thymus is the same as the recognition threshold in the periphery, i.e. $E_r = E_n$. This means that a foreign peptide of sequence $\vec{s}$ is recognized if its strongest interaction with the set of TCRs exceeds the threshold for recognition, i.e.

$$\min_{\vec{i} \in T} \{ E_{\text{int}}(\vec{i}, \vec{s}) \} < E_n. \tag{3.8}$$

Equation (3.8) casts the recognition of foreign peptides as another extreme value problem. From the previous section we already know how to calculate the probability

---

Figure 3-3: Amino-acid composition of selected TCR sequences, ordered in increasing frequency along the abscissa. The data points in black are obtained numerically with the parameters relevant to mouse (see caption of Figure 3-2, and text). The error bars reflect the sample size used to generate the histograms and differences for different realizations of $M$ self-peptides. The red line is the result of the EVD analysis in the large $N$ limit from Equation (3.7), and the agreement is quite good. The data points in green are obtained numerically with the parameters relevant to mouse, where the effect of positive selection was ignored ($((E_p - E_n)/N \rightarrow -\infty$, i.e. every TCR is positively selected). In all cases we have used the Miyazawa-Jernigan matrix $J$ (Miyazawa and Jernigan, 1996), and amino acid frequencies $f_a$ from the mouse proteome (Table 2.1 and (Flicek et al., 2008; Košmrlj et al., 2008)). This figure is adapted from Figure 3 in (Košmrlj et al., 2009).
$P_{\text{rec}}(\vec{s})$ that a foreign peptide sequence $\vec{s}$ is recognized by T cells and how to calculate the amino acid composition of recognized foreign peptides in the limit of long peptide sequences. Therefore, here we just briefly summarize the necessary steps.

Let us indicate by $\rho^*(x|\vec{s})$ the PDF of the interaction energy between the foreign peptide $\vec{s}$ and a random TCR that is selected in the thymus. Then the probability that foreign peptide $\vec{s}$ is recognized is obtained by integrating the extreme value distribution $\Pi^*(x|\vec{s})$ over the allowed range, as

$$P_{\text{rec}}(\vec{s}) = \int_{-\infty}^{E_a} \Pi^*(x|\vec{s}) \, dx, \quad \text{with}$$

$$\Pi^*(x|\vec{s}) = K \rho^*(x|\vec{s}) (1 - P^*(E < x|\vec{s}))^{K - 1}. \quad (3.9)$$

(Note that compared to Equation (3.3), there is no lower bound to the interaction strength.) If we model the set $\mathcal{T}$ of selected T cells as $K$ strings in which each amino acid is chosen independently with frequencies $f_a^{(\text{sel})}$ (i.e. we ignore the correlations among different positions on the string), then in the limit of long peptide sequences (large $N$) we can approximate the PDF $\rho^*(x|\vec{s})$ with a Gaussian with the mean $E_{\text{av}}^*(\vec{s}) = E_c + \sum_{i=1}^{N} E^*(s_i)$ and the variance $V^*(\vec{s}) = \sum_{i=1}^{N} V^*(s_i)$. The mean $E^*(s_i)$ and the variance $V^*(s_i)$ of the amino acid interaction energies are obtained as in the previous section by appropriately replacing $\{f_a\}$ with $\{f_a^{(\text{sel})}\}$. In the limit of large number of T cells ($K \gg 1$) the extreme value distribution $\Pi^*(x|\vec{s})$ is sharply peaked around

$$E_0^*(\vec{s}) = E_{\text{av}}^*(\vec{s}) - \sqrt{2V^*(\vec{s}) \ln K}, \quad (3.10)$$

and in the large $N$ limit the condition for recognition of foreign peptides becomes

$$E_0^*(\vec{s}) < E_n. \quad (3.11)$$

The probability for a sequence $\vec{s}$ to be recognized is governed by the Boltzmann weight $p(\vec{s}) \propto \left( \prod_{i=1}^{N} \tilde{f}_{s_i} \right) \exp[-\beta^* E_0^*(\vec{s})]$, where $\{\tilde{f}_a\}$ are natural frequencies of different amino acids in the pathogen proteome, while the effect of TCR recognition is captured in the parameter $\beta^*$. As in the previous section, we introduce a new Hamil-
Figure 3-4: Color representation of the dependence of the parameter $\beta^*$ on the number of selected TCRs $(\ln(K)/N)$, the number of self peptides $(\ln(M)/N)$ against which TCRs were selected, and threshold for negative selection $E_n/N$ with $(E_p - E_n)/N = 0.5k_BT$, in the limit of large $N$. Solid black lines separate regions with $\beta^* > 0$ and $\beta^* = 0$. The region between the black dashed lines corresponds to $\beta^* = 0$ (every TCR is selected). (a) $(E_n - E_c)/N = -4.2k_BT$; (b) $\ln(K)/N = 1.5$.

For a given Hamiltonian $H_0(s) = E_e + \sum_{i=1}^{N} [E^*(s_i) - \gamma^*V^*(s_i)] - \ln K/(2\gamma^*)$, and to ensure the same average energies $\langle E^*_0(s) \rangle$ and $\langle H_0^*(s) \rangle$ we set $\gamma^*(\beta^*) = \sqrt{\ln K/(2N\langle V^*(\beta^*,\gamma^*) \rangle)}$. Finally, the value of $\beta^*$ is determined by constraining the average energy $\langle E^*_0(s) \rangle < E_n$, while maximizing entropy. If $E^{*}_{\text{mid}} > E_n$, we must set $\beta^*$ such that $\langle E^*_0(s) \rangle = E_n$, where $E^{*}_{\text{mid}}$ is defined like in the previous section. In this case, $\beta^* > 0$, and only foreign peptides with stronger amino acids are recognized. If $E^{*}_{\text{mid}} < E_n$, we must set $\beta^* = 0$, and there is no modification due to recognition, i.e. every foreign peptide is recognized. Note that unlike for the thymic selection of T cell receptors (the parameter $\beta$), the parameter $\beta^*$ cannot be negative, because there is no lower energy bound for recognition in Equation (3.11). With all parameters determined, we can calculate the amino acid frequencies of recognized foreign peptides

$$f_{\text{rec}}(a) = \frac{\hat{f}_a \exp \left[ -\beta^*(E^*(a) - \gamma^*V^*(a)) \right]}{\sum_{b=1}^{20} \hat{f}_b \exp \left[ -\beta^*(E^*(b) - \gamma^*V^*(b)) \right]}.$$  

(3.12)

Figure 3-4 depicts variation of $\beta^*$ as a function of the number of selected TCRs $(\ln(K)/N)$, the number of self peptides $(\ln(M)/N)$ against which TCRs were selected,
Figure 3-5: Amino-acid composition of recognized foreign peptides, ordered in decreasing frequency along the abscissa. The data points in black are obtained numerically with the parameters relevant to the TCR selection in mouse and $K = 10^3$ TCRs where than challenged with *Listeria monocytogenes* peptides (see text). The error bars reflect the sample size used to generate the histograms and differences for different realizations of $M$ self-peptides. The red line is the result of the EVD analysis in the large $N$ limit from Equation (3.12), where TCR amino acid frequencies were obtained from Equation (3.7). The dashed lines correspond to the result of the EVD analysis in the large $N$ limit, where the selected TCR amino acid frequencies were taken from simulation results (dashed blue line) or where the effective number $K_{\text{eff}} = 0.67 \times 10^3$ of TCRs is used to take into account the correlations of selected TCR sequences (dashed green line). In all cases we have used the Miyazawa-Jernigan matrix $J$ (Miyazawa and Jernigan, 1996), amino acid frequencies $f_a$ from the mouse proteome and amino acid frequencies $\tilde{f}_a$ from the *Listeria monocytogenes* proteome (Table 2.1 and (Flicek et al., 2008; Košmrlj et al., 2008)).

and the threshold for negative selection $E_a/N$ with $(E_p - E_n)/N = 0.5k_BT$. The region $\beta^* = 0$ is possible only for $K \gtrsim M$ and for parameters where $|\beta|$ is small (cf. Figure 3-2). Finally we compared the analytical results, which are exact in the limit $N \to \infty$ with numerical simulations for $N = 5$ (Figure 3-5). From the set of parameters that are relevant for thymic selection in mouse (i.e. $N = 5$, $E_p - E_n = 2.5k_BT$, $E_n - E_c = -21k_BT$ and $M = 10^3$), we generated a pool of $K = 10^3$ selected TCRs. Then we randomly generated $10^6$ foreign peptides with amino acid frequencies $\tilde{f}_a$, which were representative of *Listeria monocytogenes* (Table 2.1), and checked the
amino acid composition of foreign peptides that were recognized by at least one TCR. We find that selected TCRs recognize only foreign peptides that are enriched with strongly interacting amino acids (Figure 3-5). The agreement between simulations (black line) and analytical result (red line, $\beta^* = 0.11(k_BT)^{-1}$ and $\gamma^* = 0.96(k_BT)^{-1}$) is not very good (Figure 3-5). One source of discrepancy could be due to differences in amino acid composition of selected TCRs $f_a^{(sel)}$ (cf. Figure 3-3). Using $f_a^{(sel)}$ obtained from simulations as an input for analytical result (dashed blue line, $\beta^* = 0.26(k_BT)^{-1}$ and $\gamma^* = 1.06(k_BT)^{-1}$) makes only slight improvement. This suggests that our assumption that the pool $T$ of selected TCRs is uncorrelated, might not be good for short sequences. To test the effect of correlations, we generated a set of $K = 10^3$ uncorrelated TCRs with amino acid frequencies $f_a^{(sel)}$ obtained from Equation (3.7) and then checked the amino acid composition of foreign peptides recognized by this set. Figure 3-6 shows better agreement between the analytical result (red line) and simulations (blue data points) with uncorrelated TCRs. However, the analytical result is still off for up to 40% and the large discrepancy is likely due to the inaccurate approximation of micro-canonical with the canonical ensemble of recognized foreign peptides for short peptides ($N = 5$). Worse agreement between the numerical and analytical results for the amino acid composition of recognized peptides (Figure 3-6) compared to the results for the amino acid composition of selected TCRs (Figure 3-3) might be due to lower magnitude of numerically obtained parameters $|\beta^*| = 0.11k_BT^{-1} < |\beta| = 0.37k_BT^{-1}$ and the exponential dependence of amino acid frequencies on parameters $\beta$ and $\beta^*$ (Equations (3.7) and (3.12)).

To quantify the degree of correlations, we calculated the "effective" TCR repertoire size from the Shannon entropy (Shannon, 1948)

$$S_N = \sum_{\{\vec{t}\}} -p_{\vec{t}}\ln p_{\vec{t}}, \quad (3.13)$$

where $p_{\vec{t}}$ is the (joint) probability for the TCR sequence $\vec{t}$ to appear in the selected TCR repertoire. It is given by $p_{\vec{t}} = C \left[ \prod_{i=1}^{N} f_{t_i} \right] \times P_{\text{sel}}(\vec{t})$, where the product of $f_{t_i}$ factors corresponds to the probability that the TCR sequence $\vec{t}$ is generated, the
Figure 3-6: Amino-acid composition of foreign peptides recognized by uncorrelated TCRs, ordered in decreasing frequency along the abscissa. The data points in blue are obtained numerically where a set of \( K = 10^3 \) uncorrelated TCRs is challenged with *Listeria monocytogenes* peptides (see text). The error bars reflect the sample size used to generate the histograms and differences for different realizations of \( K \) uncorrelated TCRs. The red line is the result of the EVD analysis in the large \( N \) limit from Equation (3.12), where TCR amino acid frequencies were obtained from Equation (3.7). In both cases we have used the Miyazawa-Jernigan matrix \( J \) (Miyazawa and Jernigan, 1996), amino acid frequencies \( f_a^{(\text{sel})} \) from Equation (3.7) and amino acid frequencies \( f_a \) from the *Listeria monocytogenes* proteome (Table 2.1 and (Flicek et al., 2008; Košmrlj et al., 2008)).

\( P_{\text{sel}}(\tilde{t}) \) corresponds to the probability that the TCR sequence \( \tilde{t} \) is selected in the thymus, and \( C \) is the normalization constant. For the model parameters relevant to mouse (\( N = 5, E_n - E_c = -21k_B T, E_p - E_n = 2.5k_B T \) and \( M = 10^3 \)), we numerically estimated the probabilities \( P_{\text{sel}}(\tilde{t}) \) for each of the \( 20^N \) sequences \( \tilde{t} \), by calculating the average probability that a given sequence \( t \) is selected against \( M = 10^3 \) self peptides averaged over 1,000 different realizations of \( M \) self peptides. Thus we obtained numerically the “effective” TCR repertoire size \( \exp(S_N) \approx 0.64 \times 10^6 \). Assuming that the selected TCR sequences are uncorrelated, the “effective” TCR repertoire size can be calculated also from the amino acid frequencies \( f_a^{(\text{sel})} \) of selected
TCRs. In this case the Shannon entropy is

\[ S_1 = \sum_{a=1}^{20} -f_a^{(sel)} \ln f_a^{(sel)}. \] (3.14)

For the amino acid frequencies \( f_a^{(sel)} \) obtained from numerical simulations, the size of the “effective” TCR repertoire is \( \exp(NS_1) \approx 0.96 \times 10^6 \). The size of the “effective” TCR repertoire obtained for analytically determined amino acid frequencies \( f_a^{(sel)} \) from Equation (3.7) is \( \exp(NS_1) \approx 1.08 \times 10^6 \). The correlations thus reduce the “effective” TCR repertoire size by approximately one third. In analytical results (Equation (3.12)) the effect of correlations can be approximated by using an effective number of TCRs \( K_{\text{eff}} \approx \frac{\exp(S_N)}{\exp(NS_1)} K \). Using \( K_{\text{eff}} \approx 0.67 \times 10^3 \) leads to parameters \( \beta^* = 0.14(k_B T)^{-1} \) and \( \gamma^* = 0.93(k_B T)^{-1} \), resulting only in a very slight change of amino acid frequencies of recognized peptides (dashed green line in Figure 3-5) and they are still far from the results of simulations (black data points). This suggests that the effect of correlations is much larger for short sequences \( (N = 5) \) and for the chosen parameters.

It is important to note that a mouse has \( \sim 10^8 \) distinct T cells and a human has \( \sim 10^9 \) distinct T cells. Both numbers are larger than the size of the “effective” TCR repertoire \( (\sim 10^6) \), as well as the total number of possible sequences \( 20^N \) for \( N = 5 \). This suggests that the sequence length \( N \) should be larger \( (N = 6 \text{ or } N = 7) \) or that TCRs differ also in regions that do not bind peptide, e.g. there are different sets of TCR pools that correspond to different MHC types (each human has up to 12 different MHC types). For the values of \( K \) that are of the order of the “effective” repertoire size or larger, the EVD \( \Pi^* \) probes the tails of distribution \( \rho^*(x|\vec{s}) \), where it is no longer Gaussian. Because the distribution \( \rho^*(x|\vec{s}) \) is bounded, in the large \( K \) limit the EVD approaches a delta-function centered at the \( K \)-independent value corresponding to the optimal binding energy:

\[ E_0^*(\vec{s}) = E_c + \min_{\{\tilde{r} \in T\}} \sum_{i=1}^{N} J(t_i, s_i). \] (3.15)

70
In the limit $N \to \infty$, a pool of TCR sequences is uncorrelated and the optimal binding energy can be written as:

$$E_0^*(\tilde{s}) = E_c + \sum_{i=1}^{N} J_{\text{min}}(s_i), \quad (3.16)$$

where $J_{\text{min}}(a) = \min_b J(b, a)$. The condition for recognition of foreign peptides is

$$E_0^*(\tilde{s}) < E_n, \quad (3.17)$$

and the probability for a sequence $\tilde{s}$ to be recognized is governed by the Boltzmann weight $p(\tilde{s}) \propto \left( \prod_{i=1}^{N} \tilde{f}_{s_i} \right) \exp[-\beta^*E_0^*(\tilde{s})]$, where $\{\tilde{f}_{a}\}$ are natural frequencies of different amino acids in the pathogen proteome, while the effect of TCR recognition is captured in the parameter $\beta^*$. The value of $\beta^*$ is determined by constraining the average energy $\langle E_0^*(\tilde{s}) \rangle < E_n$, while maximizing entropy in the same manner as described before (see the paragraph after Equation (3.11)). The condition for $\beta^* > 0$ can be simplified to:

$$\frac{E_n - E_c}{N} = \langle J_{\text{min}}(a) \rangle = \frac{\sum_{a=1}^{20} J_{\text{min}}(a) \tilde{f}_{a} \exp[-\beta^*J_{\text{min}}(a)]}{\sum_{a=1}^{20} \tilde{f}_{a} \exp[-\beta^*J_{\text{min}}(a)]}. \quad (3.18)$$

With the parameter $\beta^*$ determined, we can calculate the amino acid frequencies of recognized foreign peptides

$$\tilde{f}_{a}^{(\text{rec})} = \frac{\tilde{f}_{a} \exp[-\beta^*J_{\text{min}}(a)]}{\sum_{b=1}^{20} \tilde{f}_{b} \exp[-\beta^*J_{\text{min}}(b)]}. \quad (3.19)$$

For the relevant parameters in mouse (i.e. $N = 5$, $E_p - E_n = 2.5k_BT$, $E_n - E_c = -21k_BT$ and $M = 10^3$), we obtain $\beta^* = 0$, which means that every peptide is recognized. We tested this numerically by generating all $20^N$ possible TCR sequences $\tilde{r}$ with appropriate weights ($\prod_{i=1}^{N} f_{t_i}$), which are then selected in the thymus against $M = 10^3$ self peptides and then challenged against $10^5$ randomly generated foreign peptides that were representative of *Listeria monocytogenes*. Figure 3-7 shows a large disagreement between numerical simulations (black data points) and this analytical
Figure 3-7: Amino-acid composition of foreign peptides recognized by a complete pool of selected TCRs, ordered in decreasing frequency along the abscissa. The data points in black are obtained numerically where a complete pool of selected TCRs (obtained from thymic selection of all all $20^N$ TCR sequences against $M = 10^3$ self peptides) is challenged with $L$. monocytogenes peptides (see text). The error bars reflect the differences for different realizations of $M = 10^3$ self peptides. The red line is the result of the EVD analysis in the large $N$ limit from Equation (3.19). The data points in blue are obtained numerically where a complete pool of uncorrelated TCRs, which is equal to the complete pool of $20^N$ sequences (see text), is challenged with $L$. monocytogenes peptides. In this case, there are no errorbars, because we use the complete pools of TCRs and foreign peptides with appropriate weights. In all cases we have used the Miyazawa-Jernigan matrix $J$ (Miyazawa and Jernigan, 1996), amino acid frequencies $f_a$ from human and amino acid frequencies $\tilde{f}_a$ from the $L$. monocytogenes proteome (Table 2.1 and (Flicek et al., 2008; Košmrlj et al., 2008)).

result (red line). Because the selected TCRs are correlated, we also tested the effect of correlations. Since $f_a^{(sel)} > 0$ for every amino acid $a$, the complete pool of uncorrelated TCRs correspond to all $20^N$ sequences with weights $(\prod_{i=1}^N f_{t_i}^{(sel)})$. When a foreign peptide is tested against the complete pool of TCRs, it finds a TCR that results in optimal binding energy (Equation 3.16). To numerically calculate the amino acid frequencies of foreign proteins recognized by complete uncorrelated pool of TCRs, we generate all possible $20^N$ foreign peptide sequences with appropriate weights $(\prod_{i=1}^N \tilde{f}_{s_i})$. A foreign peptide sequence $\tilde{s}$ is then recognized if the optimal binding energy (Equation 3.16)
exceeds the recognition threshold (Equation 3.17). The numerical result for uncorrelated TCRs (blue data points) agrees very well with the analytical results, which indicates that the correlations in selected TCR sequences have an important role.

3.3 Conclusions

Equation (3.7) provides an analytical expression that captures the characteristics of TCR amino-acids selected against many peptides in the thymus. In accord with previous numerical results presented in Chapter 2 and reference (Košmrlj et al., 2008), and some available data from normal mouse and human, it predicts (since $\beta < 0$) that TCR sequences are enriched in weakly interacting amino-acids (small $\mathcal{E}$). This result was used previously (Košmrlj et al., 2008) to explain their specificity. However, Equation (3.7) further indicates the role of promiscuity of amino-acids (captured by the parameter $\gamma$) which was not elucidated from the limited numerical data. Furthermore, the phase diagram in Figure 3-2 indicates how upon raising the number of self-peptides there is a transition from preference for strong amino-acids ($\beta > 0$, positive selection dominant) to weak amino-acids ($\beta < 0$, negative selection dominant), which may be feasibly tested in future experiments, along the lines in reference (Huseby et al., 2006).

Equations (3.12) and (3.19) provide an analytical expression that captures the characteristics of amino-acids of the recognized foreign peptides. Figures 3-5 and 3-7 show that the analytical result is not very accurate for short ($N = 5$) peptide sequences and we showed (Figures 3-6 and 3-7) that this is due to the correlations in selected TCR sequences. In the future it would be interesting to study how (or if) these correlations vanish as peptide sequence length ($N$) is increased.
Chapter 4

Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection

Without therapy, most people infected with human immunodeficiency virus (HIV) ultimately progress to AIDS. Rare individuals (elite controllers) maintain very low levels of HIV RNA without therapy, thereby making disease progression and transmission unlikely. Certain HLA class I alleles (human MHC is called HLA) are markedly enriched in elite controllers, with the highest association observed for HLA-B57 (Migueles et al., 2000). Because HLA molecules present viral peptides that activate CD8+ T cells, an immune-mediated mechanism is probably responsible for superior control of HIV. In this Chapter, we describe how the peptide-binding characteristics of HLA-B57 molecules affect thymic development such that, compared to other HLA-restricted T cells, a larger fraction of the naive repertoire of B57-restricted clones recognizes a viral epitope, and these T cells are more cross-reactive to mutants of targeted epitopes. Our calculations predict that such a T-cell repertoire imposes strong

---

This chapter is based on the publication (Košmrlj et al., 2010).
immune pressure on immunodominant HIV epitopes and emergent mutants, thereby promoting efficient control of the virus. Supporting these predictions, in a large cohort of HLA-typed individuals, our experiments show that the relative ability of HLA-B alleles to control HIV correlates with their peptide-binding characteristics that affect thymic development. Our results provide a conceptual framework that unifies diverse empirical observations, and have implications for vaccination strategies.

HIV infection leads to acute high level viraemia (the measurable presence of virus in the bloodstream), which is subsequently reduced to a set-point viral load. Without therapy, most patients experience a subsequent increase in viral load, and ultimately the development of AIDS. Viraemia levels and time to disease vary widely, and the differences correlate with the expression of different HLA class I molecules (reviewed in reference (Deeks and Walker, 2007)). Effector CD8+ T cells (CTLs) are implicated in viral control because T-cell antigen receptors (TCRs) on CD8+ T cells recognize complexes of viral peptides and class I HLA molecules presented on the surface of infected cells, and depletion of CD8+ T cells leads to increased viraemia in animal models of HIV infection (Jin et al., 1999). We describe a feature of the HLA-B57-restricted CD8+ T-cell repertoire that contributes to enhanced control of viraemia.

4.1 HLA-B57 binds fewer types of self peptides

Algorithms (Peters et al., 2005) based on experimental data predict whether a particular peptide will bind to a given HLA molecule (Rao et al., 2009). We tested four predictive algorithms against available experimental data on peptide binding to diverse HLA molecules and found that, in most cases, they are highly accurate (Figure B-1 and Table B.1). For example, predictions using the best algorithm for HLA-B*5701 were 97% accurate. Using these algorithms, we computed the fraction of peptides derived from the human proteome (Hubbard et al., 2009) that bind to various HLA molecules. Of the $\sim 10^7$ unique peptide sequences, only 70,000 are predicted to bind to HLA-B*5701, and 180,000 bind to HLA-B*0701 (an allele that is not protective against HIV). Essentially identical results were obtained for randomly
generated peptides (data not shown). The protective allele in macaques, Mamu-B*17, also binds fewer self peptides than other Mamu molecules for which data are available (Mamu-B*17 binds 4, 6 and 13 times fewer self peptides than Mamu-A*11, Mamu-A*01 and Mamu-A*02, respectively; Table B.1).

The intrinsic differences in self-peptide binding among HLA molecules are important during T-cell repertoire development. Immature T cells are exposed to diverse host-derived peptide-HLA complexes presented in the thymus. As fewer self peptides are able to bind to HLA-B*5701 (and Mamu-B*17) molecules, a smaller diversity of self-peptide TCR contact sequences will be encountered by HLA-B*5701/Mamu-B*17-restricted T cells in the thymus.

### 4.2 Thymic selection against fewer types of self peptides results in a more cross-reactive T-cell repertoire

The diversity of self peptides presented in the thymus shapes the characteristics of the mature T-cell repertoire. Experiments (Huseby et al., 2006, 2005) and theoretical studies (Chapters 2 and 3 and (Kosmrlj et al., 2008, 2009)) show that T cells that develop in mice with only one type of peptide in the thymus are more cross-reactive to point mutants of peptide epitopes that they recognize than T cells from mice that express diverse self peptides. Thus, by encountering fewer self peptides during thymic development, HLA-B57-restricted CD8+ T cells should be more cross-reactive to point mutants of targeted viral peptides.

Our study showed that the protective allele HLA-B*5701 binds fewer peptides derived from human proteome compared to other alleles (Table B.1). Even if the reason why HLA-B*5701 molecules bind fewer self peptides was due to greater restrictions in the tolerance to different amino acids at the anchor residues only, HLA-B*5701 molecules would present a smaller diversity of TCR contact residues in the thymus. This is because the number of self peptides presented in the thymus is much smaller than all possible sequences of TCR contact residues derived from the human proteome. Thus, the probability that any HLA allele presents peptides derived from different parts of the proteome with identical TCR contact residues constrained by the same anchor residues is small. Therefore, since HLA-B*5701 presents fewer self-peptides, T cells restricted by this allele will encounter a smaller diversity of TCR contact residues during development in the thymus.
We carried out in silico thymic selection experiments to test this hypothesis. We chose an HLA-dependent number of thymic self peptides, each with amino acids of the TCR contact residues picked according to the frequency with which they appear in the human proteome (Table 2.1 and (Hubbard et al., 2009; Košmrlj et al., 2008)). A diverse set of immature CD8$^+$ T cells (thymocytes) was generated by choosing the sequences of their peptide contact residues in the same way, and by varying the TCR–HLA interactions. A thymocyte emerges from the thymus as a mature CD8$^+$ T cell if its TCR binds to at least one self-peptide–major histocompatibility complex (pMHC; human MHC is called HLA) molecule with an affinity that exceeds the positive selection threshold, and does not interact with any pMHC more strongly than the negative selection threshold. Using a computational model (Chapters 2 and 3 and (Košmrlj et al., 2008, 2009)) in the class of string models (Chao et al., 2005), we assessed the affinity of TCR–self-peptide–HLA complexes to determine which T cells survive positive and negative selection, and become a part of the mature repertoire. Our qualitative results are independent of the parameters used to determine these interaction strengths (see Chapters 2 and 3, Appendix A and (Košmrlj et al., 2008, 2009)).

The mature T cells that emerged from these in silico thymic selection experiments were then computationally challenged by a viral peptide (that is, not seen in the thymus) bound to the same HLA type. T cells that recognize this peptide–HLA complex were obtained by assessing whether the interaction strength exceeded the negative selection threshold (shown to be equal to the recognition threshold in mouse models (Naeher et al., 2007)). Cross-reactivity of these T cells was then determined in silico by mutating each TCR contact residue of the peptide to the other 19 possibilities. Sites on the viral peptide were called important contacts if half the mutations therein abrogated recognition by T cells that target this epitope. The frequency of the number of important contacts in viral peptides that determine T-cell recognition was obtained by repeating this procedure 1,000 times with different choices of thymocytes and self and foreign peptides.

Our calculations predict that a T-cell repertoire restricted by an HLA molecule
such as HLA-B*5701, which presents fewer self peptides in the thymus, has a higher frequency of occurrence of T cells that recognize viral peptides through smaller numbers of important contacts (Figure 4-1a). In contrast, the frequency of occurrence of T cells that recognize viral peptides through many important contacts is larger for repertoires restricted by HLA alleles that present a greater diversity of self peptides in the thymus (data not shown for ≥4 contacts). Mutations at sites different from the important contacts do not affect binding strength substantially. Therefore, when the interaction between peptide–HLA and TCR is mediated by fewer important contacts, a larger number of possible point mutations of the peptide do not affect peptide recognition, thereby making the T cells more cross-reactive to mutants that arise. Thus, the HLA-B57-restricted T-cell repertoire is expected to be more cross-reactive to mutants of targeted viral peptides than repertoires restricted by HLA alleles that present a greater diversity of self peptides.

Our computational models give this qualitative mechanistic insight, but do not provide quantitative estimates of the extent of this enhanced cross-reactivity of T cells. However, compelling experimental data (Turnbull et al., 2006) has shown that the effect revealed by our studies is important in humans. Peripheral blood mononuclear cells from patients expressing HLA-B57 contained CTLs that were more cross-reactive to various HIV epitopes and their point mutants than those of HLA-B8-positive patients. HLA-B8 is associated with rapid progression to disease (Turnbull et al., 2006), and the most accurate algorithm for peptide binding suggests that the HLA-B8 molecule binds a greater diversity of self peptides than HLA-B57 (Figure B-2 and Table B.1). Other experimental studies also show that patients expressing HLA-B57 cross-recognize point mutants of the dominant epitope and use more public TCRs (i.e. cross-reactive TCRs present in more than one person; (Gillespie et al., 2002; Yu et al., 2007)).

Next, we computed interaction strengths between diverse viral peptides and members of T-cell repertoires restricted by HLA molecules that present differing numbers of self peptides in the thymus. This allowed us to obtain the probability with which a randomly picked T-cell clone and viral peptide will interact sufficiently strongly for
Figure 4-1: Thymic selection against fewer self peptides leads to a more cross-reactive T-cell repertoire. (a) Histogram of the frequency with which T cells recognize viral peptides (that is, not seen in the thymus) through only a small number (0, 1, 2, 3) of important contacts (see text) is shown for three T-cell repertoires that developed with different numbers of self-peptide–HLA complexes in the thymus. If the TCR–peptide–HLA interaction is sufficiently strong, no single point mutation can abrogate recognition, resulting in zero important contacts. A higher frequency of occurrence of a small number of important contacts indicates a more cross-reactive T-cell repertoire because only mutations at these contacts are likely to abrogate recognition. (b) The probability that a TCR binds to viral peptides with a certain interaction strength is shown for three T-cell repertoires. A particular TCR recognizes a viral peptide when the binding strength exceeds the recognition threshold (dashed black line). Members of a T-cell repertoire selected against fewer self peptides are more likely to recognize a viral peptide. The model we used describes qualitative trends robustly (Chapters 2 and 3 and (Kosmrlj et al., 2008, 2009)), but is not meant to be quantitatively accurate. This figure is adapted from Figure 1 in (Kosmrlj et al., 2010).

recognition to occur. The results (Figure 4-1b) indicate that a typical CD8+ T cell restricted by an HLA molecule such as HLA-B*5701, which presents fewer peptides in the thymus, has a higher probability of recognizing a viral epitope compared to a T cell restricted by other HLA molecules. Thus, more HLA-B*5701-restricted T cell clones are likely to recognize a viral epitope, making effective precursor frequencies higher in an HLA-B*5701-restricted repertoire (a strong predictor of response magnitude (Moon et al., 2007)). A greater precursor frequency for viral epitopes in the naive repertoire restricted by HLA-B57 is indicated by experimental results showing that HLA-B*5701 contributes the most to acute-phase CTL responses of all HLA alleles tested (Altfeld et al., 2006).
The results in Figure 4-1 stem from the constraint that thymocytes must avoid being negatively selected by each self-peptide–HLA complex encountered during development in the thymus. T cells expressing TCRs with peptide contact residues composed of amino acids that interact strongly with other amino acids (for example, charged residues, flexible side chains) have a high probability of binding to a self peptide strongly. The greater the diversity of self peptides presented in the thymus, the higher the chance that a TCR with such peptide contact residues will encounter a self peptide with which strong interactions will result in negative selection. Thus, as the diversity of self peptides presented in the thymus increases, the peptide contact residues of TCRs in the mature T-cell repertoire are increasingly enriched in weakly interacting amino acids (shown analytically in Figure 3-2 and numerically in Figure A-8). T cells bearing TCRs with weakly interacting peptide contact residues recognize viral peptides by means of several moderate interactions, making many contacts important for recognition. In contrast, TCRs with peptide contact residues containing strongly interacting amino acids are more likely to recognize viral peptides through a few important contacts mediated by these residues, making recognition cross-reactive to mutations at other peptide sites. These mechanistic insights are supported by experimental results (see Chapter 2 and (Huseby et al., 2006; Košmrlj et al., 2008)).

4.3 Cross-reactivity of T cell repertoire correlates with the control of HIV infection.

By studying a model of host–pathogen dynamics\(^2\) that builds on past models of host–HIV interactions (Althaus and de Boer, 2008; Nowak et al., 1995; Wodarz and Thomsen, 2005), we explored the consequences of the HLA-B57-restricted CD8\(^+\) T-cell repertoire having a higher precursor frequency for viral peptides, and being more cross-reactive to point mutants of targeted epitopes on the control of HIV. Because of the importance of immune control exerted by CD8\(^+\) T cells (Altfeld et al., 2006; Cao et al., 2003), we focused on the interaction between a mutating virus quasispecies and

\(^2\)The model of host–pathogen dynamics was developed by Elizabeth Read.
Figure 4-2: Dynamic model of host–pathogen interactions: the virus mutates, infects limited-target CD4+ T cells, and is cleared. Infected CD4+ T cells produce more free virus and die. Infected cells present viral peptides in complex with HLA molecules (until peptides unbind from HLA). Activated CD8+ T cells produced by recognition of viral epitopes on antigen-presenting cells (APCs) proliferate and differentiate into effector CTLs. CTLs kill infected cells bearing cognate peptide–HLA complexes, and turn into memory cells that are activated after re-exposure to antigen. This figure is adapted from Figure 2a in (Kosmrlj et al., 2010).

epitope-directed, variably cross-reactive, host CTL responses.

The essential features of the model are depicted in Figure 4-2 (details in Appendix C). The virus is modelled as a number of epitopes consisting of strings of amino acids, and new viral strains (point mutations of epitopes), which differ in replicative fitness, arise over the course of infection. For each individual, an HLA-dependent CD8+ T-cell repertoire was chosen. To mimic the results obtained from our thymic selection calculations (Figure 4-1b), more or less cross-reactive repertoires were chosen (Figure C-1) to represent HLA-B57-restricted T cells and those restricted by other HLAs, respectively. Infection rates were limited by target CD4+ T cells, and CTL contraction and memory were included. Other dynamic models were studied, including one that does not incorporate target cell limitation or CTL contraction. Our qualitative results about the effects of cross-reactivity are robust to variations in
Figure 4-3: Simulated HIV viral loads versus time for different cross-reactivities (CR) of the CD8+ T-cell repertoire (the black and red curves correspond to high and low cross-reactivity respectively). Each curve is averaged over 500 simulations (each simulation represents a person). The model shows a reduced set-point viral load for people with a more cross-reactive T-cell repertoire. Other models of host–pathogen dynamics show similar effects of T-cell cross-reactivity (Figures C-2 and C-3). This figure is adapted from Figure 2b in (Kosmrlj et al., 2010).

parameters and model assumptions (Figures C-2 – C-11).

We find that individuals with a more cross-reactive CTL repertoire control viral loads better during the acute phase of the infection (Figure 4-3). This is in agreement with findings in simian immunodeficiency virus (SIV)-infected rhesus macaques (Price et al., 2009), where the number of cross-reactive TCR clones negatively correlates with viral load. Our simulations show that a larger number of CTL clones in a more cross-reactive T-cell repertoire recognize epitopes from the infecting viral strain (Figure 4-4). This is because the predicted higher precursor frequency for viral epitopes (Figure 4-1b) leads to a greater response magnitude (as in mouse models (Moon et al., 2007)). This conclusion is supported by data showing that in people with a protective HLA allele, the initial T-cell response to HIV is dominated by T cells restricted by the protective HLA and not those restricted by other HLAs expressed (Altfeld et al., 2006). Our simulations also show that enhanced cross-reactivity of the T-cell repertoire leads to greater immune pressure on the emergent viral mutants by individuals
expressing HLA-B57 compared to those with T cells restricted by HLA molecules that bind more types of self peptides. The stronger immune pressure on infecting and emerging viral strains results in superior control of viral load. Thus, we predict that HIV-infected individuals with HLA alleles that bind fewer self peptides are more likely to control viral loads to low values.

To test this prediction, we studied two large HLA-typed cohorts: 1,110 controllers with less than 2,000 HIV particles ml\(^{-1}\) and 628 progressors (or non-controllers) with viral loads exceeding \(10^4\) ml\(^{-1}\). From these data, we obtained the odds ratio (OR)
for individual HLA alleles. The odds ratio (OR) for an allele is defined as: 
\[
\frac{p_w}{p_{wo}} \frac{c_w}{c_{wo}},
\]
where \(p_w\) and \(p_{wo}\) are the numbers of individuals in the progressor cohort with and without this HLA, respectively; and \(c_w\) and \(c_{wo}\) are the numbers of individuals in the controller cohort with and without this HLA, respectively. This definition suggests that the OR measures the likelihood of an allele being correlated with progressors versus controllers, with an OR greater (less) than one indicating association with the progressor (controller) cohort. We focused on HLA-B alleles because they are associated with control of HIV (Kiepiela et al., 2004). Of 40 HLA-B alleles that were studied, significant results (P value < 0.05)\(^3\) were obtained for five HLA-B alleles (Table 4.1) and peptide-binding data are available for four of them. In support of our predictions, those HLA-B alleles associated with higher OR values also bind more self peptides (Figure 4-5).

Superior control of viral load due to the greater precursor frequency and cross-reactivity of those T-cell repertoires restricted by HLA molecules that bind to few self peptides (for example, HLA-B57) should also confer protection against diseases caused by other fast-mutating viruses. Indeed, HLA-B57 is protective against hepatitis C

\(^3\)HLA-allele association with ability to control HIV. SAS 9.1 (SAS Institute) was used for data management and statistical analyses. Odds ratios and 95% confidence intervals were determined using PROC LOGISTIC in a comparison of HIV controllers (those individuals who maintained viral loads of less than 2,000 copies of the virus per ml plasma on three determinations over at least a year of follow-up and, on average, for approximately 15 years (Pereyra et al., 2008)) to HIV non-controllers (those individuals whose viral loads exceeded 10,000 copies of the virus per ml plasma). To eliminate the confounding effects of HLA-B*0702, HLA-B*3501, HLA-B*2705 and HLA-B*5701, alleles strongly associated with progression or control, these factors were used as covariates in the logistic regression model for the analysis of all other HLA class I types (Hosmer et al., 1989). All ethnic groups were included in the analyses shown (European, African-American and others) and we adjusted for ethnicity in the logistical regression model. All P values were corrected for multiple tests using the Bonferroni correction, a stringent and commonly used approach for multiple comparisons (Cheverud, 2001).
Figure 4-5: HLA-B alleles associated with greater ability to control HIV correlate with smaller self-peptide binding propensities. This definition suggests that the OR measures the likelihood of an allele being correlated with progressors versus controllers, with an OR greater than one indicating association with the progressor cohort. People with HLA alleles associated with an OR value greater or less than one are more likely to be progressors or controllers, respectively. The fraction of peptides derived from the human proteome that bind to a given HLA allele was determined using the most accurate predictive algorithms (Table B.1). Compared to experimental data, the predictive algorithms for peptide binding by HLA-B*3501 are less accurate than algorithms for the other three alleles (Figure B-3 and Table B.1); the number reported here for HLA-B*3501 using the most accurate algorithm underestimates the binding fraction. The error bars represent the 95% confidence intervals for OR. The dotted line corresponds to equal odds for an allele being associated with progressors and controllers. This figure is adapted from Figure 3 in (Kosmrlj et al., 2010).

virus, HCV (Thio et al., 2002), another highly mutable viral disease in which CD8+ T cells are important. Also, HLA-B8, which binds a greater diversity of self peptides, is associated with faster disease progression in HCV (McKiernan et al., 2004) and HIV (Turnbull et al., 2006). Thus, the correlation between the diversity of peptides presented in the thymus during T-cell development and control or progression of disease may be general.
4.4 Conclusions

Undoubtedly, many complex factors influence the relationship between HLA type and disease outcome. The effect of the new factor we have identified should be greatest for HLA molecules that bind relatively few (for example, HLA-B57) or many (for example, HLA-B7, -B35, -B8) self peptides. The strong association of HLA-B27, which binds an intermediate number of self peptides (twice as many as HLA-B57), with viral control indicates that, in this case, the effects of T cell cross-reactivity are reinforced by this molecule binding HIV epitopes that are subject to very strong structural constraints.

Our results also point to a mechanistic explanation for as yet unexplained associations between HLA alleles that confer protection against HIV and autoimmune diseases. T cells restricted by HLA alleles that bind to few self peptides are subject to less stringent negative selection in the thymus, and should therefore be more prone to recognizing self peptides. Indeed, HLA-B57 has been associated with autoimmune psoriasis (Bhalerao and Bowcock, 1998) and hypersensitivity reactions (Chessman et al., 2008). Enhanced cross-reactivity of HLA-B27-restricted T cells and other unique properties of this molecule (misfolding, homodimers (de Castro, 2007)) probably contribute to the enhanced risk of autoimmunity associated with this allele (Bowness, 2002).

Our results shed light on another intriguing observation; acutely infected patients with low viral loads (and protective HLAs) tend to target an immunodominant epitope that makes a larger relative contribution to the total CTL response as compared to individuals presenting with higher levels of viraemia (Streeck et al., 2009). This is counterintuitive as the most protective responses appear most focused, rather than broadly distributed over many epitopes. We calculated how viral load correlates with the number of CTLs responding to the immunodominant epitope divided by the total number of CTLs activated by the virus (a quantity analogous to relative contribution (Streeck et al., 2009)). Mirroring experimental data, HLA alleles that restrict a more cross-reactive repertoire and are more protective also make a larger
relative contribution (Figure C-8). This result unifies the idea of both a broad and a focused response. The more cross-reactive repertoire targets more epitopes and emergent mutants, but a larger number of clones also recognize the dominant epitope (Figure 4-4).

Cross-reactive T cells are rare in people with HLA alleles that present more self peptides in the thymus than the B57 allele, but they do exist. Our results suggest that a T-cell vaccine for a diverse population must aim to activate these rare cross-reactive T cells that also target epitopes from a conserved region of the HIV genome (like HLA-B57 Gag epitopes). This will enable robust responses to infecting and mutant strains until a strain with low replicative fitness emerges, enhancing control of viral load.
Chapter 5

Thymic selection of T cells as diffusion with intermittent traps

After bacterial or viral infection, the job of an organism’s immune system is to rec-
ognize and eliminate pathogens from the host. T cells orchestrate adaptive immune
responses by recognizing short peptides derived from pathogen proteins, and by dis-
tinguishing them from self-peptides derived from host proteins (Janeway et al., 2004).
To ensure the latter, immature T cells (thymocytes) diffuse around the thymus gland,
where they encounter an ensemble of self-peptides presented on (immobile) antigen
presenting cells (APC) (Hogquist et al., 2005; Palmer, 2003). If a thymocyte binds
strongly to such an APC, it is eliminated; i.e. the APC acts as a trap for the diffusing
thymocyte. Since the peptides presented by APCs are recycled, the traps are not per-
manent, but intermittently turn “on” and “off”. Only a few percent of thymocytes
survive thymic selection (van Meerwijk et al., 1997; Merkenschlager et al., 1997), but
these few may cause autoimmune diseases. We are interested in the escape probability
of such “autoimmune” T cells during thymic selection. To address this question, in
this Chapter we study the survival probability of a random walker on a d-dimensional
cubic lattice with randomly placed immobile traps that intermittently switch between
“on” and “off” states. First we review what is known about the survival probability
in a well studied problem where traps are always “on”. We then discuss the effect of

This chapter is based on the publication (Košmrlj, 2011).
intermittent two-state traps on the survival probability at short and long times and describe a peculiar caging effect at intermediate times when switching between trap states is slow.

5.1 Classical trapping problem

The problem of trapping in a random medium has been studied extensively, with applications including dynamical processes in disordered media, kinetics of reactions, electron-hole recombinations in random and amorphous solids, and exciton trapping and annihilation (Havlin and Ben-Avraham, 1987; den Hollander and Weiss, 1994). Here we review what is known about the survival probability of a random walker in a field of randomly placed immobile traps on a $d$-dimensional cubic lattice, when averaged over all random trap configurations. At each step a walker hops from one site to a randomly chosen nearest neighbor site and if it lands on a trap, it dies. Suppose that a walker has survived the first $n$ steps, during which it has visited $S_n$ distinct lattice sites. If the trap concentration is $c$, then the probability that all $S_n$ lattice sites were not traps is equal to $(1 - c)^{S_n}$. The average survival probability of a walker after $n$ steps is thus

$$
\Phi_{ON}(n, c) = \langle (1 - c)^{S_n} \rangle = \langle e^{-\lambda S_n} \rangle,
$$

(5.1)

where $\lambda = -\ln(1 - c)$, and the averaging is over all realizations of random walks of length $n$. The exact analytical result of survival probability is only known in one dimension (den Hollander and Weiss, 1994), while analytical approximations in higher dimensions are only known for small and large $n$.

At small and intermediate number of steps $n$, the survival probability is accurately described by the Rosenstock approximation (Rosenstock, 1970)

$$
\Phi_{ON}(n, c) \approx e^{-\lambda \langle S_n \rangle}.
$$

(5.2)

Asymptotic values of $\langle S_n \rangle$ for large $n$ are known in all dimensions $d$ (Montroll and
Weiss, 1965):

\[
\begin{array}{c|c}
\hline
 d & \langle S_n \rangle \\
\hline
 1 & \sqrt{8n/\pi} \\
 2 & \pi n / \ln(8n) \\
 \geq 3 & n / P(0; 1), \\
\end{array}
\]

(5.3)

where \( P(x; z) = \sum_{n=0}^{\infty} z^n P_n(x) \) is the generating function of the probability distribution \( P_n(x) \) that a random walker is at location \( x \) after \( n \) steps. The Rosenstock approximation can be extended by rewriting the survival probability in terms of cumulants (Zumofen and Blumen, 1982):

\[
\ln \Phi_{ON} = \ln (e^{-\lambda S_n}) = \sum_{j=1}^{\infty} (-1)^j \frac{\lambda^j}{j!} \kappa_j(n),
\]

(5.4)

where the first two cumulants are \( \kappa_1(n) = \langle S_n \rangle \) and \( \kappa_2(n) = \langle S_n^2 \rangle_c = \langle (S_n - \langle S_n \rangle)^2 \rangle \). For fixed \( n \) the series converges rapidly when \( \lambda \) is small. This is because \( 1 \leq S_n \leq n + 1 \), which implies that the \( j \)th cumulant, \( \kappa_j(n) \) is at most of order \( n^j \). The series is thus expected to converge rapidly when \( n \ll \lambda^{-1} \). In practice only the first two terms can be used, because in \( d \geq 2 \) analytical expressions are known only for the first two cumulants (Montroll and Weiss, 1965; Torney, 1986). To estimate the range of validity of Rosenstock approximation, we check where \( \lambda^2 \kappa_2(n) \ll \lambda \kappa_1(n) \):

\[
\begin{array}{c|c|c|c}
\hline
 d & \kappa_1(n) & \kappa_2(n) & \lambda^2 \kappa_2(n) \ll \lambda \kappa_1(n) \\
\hline
 1 & a_1 \sqrt{n} & b_1 n & n \ll \left( \frac{a_1}{b_1 \lambda} \right)^2 \\
 2 & a_2 \frac{n}{\ln(8n)} & b_2 \frac{n^2}{\ln^2(8n)} & n \ll \frac{a_2}{b_2 \lambda} \ln^3 \left( \frac{1}{\lambda} \right) \\
 3 & a_3 n & b_3 n \ln n & n \ll \exp \left( \frac{a_3}{b_3 \lambda} \right) \\
 \geq 4 & a_d n & b_d n & \text{all } n. \\
\end{array}
\]

(5.5)

Note that the estimate above appears to suggest that for \( d \geq 4 \) the Rosenstock approximation is valid for all \( n \), but this is not correct as will be described next. Therefore it is only safe to assume that the Rosenstock approximation is valid for \( n \ll \lambda^{-1} \).
In the limit of large $n$, Donsker and Varadhan (Donsker and Varadhan, 1975, 1979) proved that the survival probability falls asymptotically as a stretched exponential:

$$\Phi_{ON}(n, c) = e^{-k_d \lambda^{2/(d+2)} n^{d/(d+2)}}, \quad (5.6)$$

where $k_d$ is numerical constant which depends on dimension $d$ and lattice properties. For a $d$-dimensional cubic lattice the value of $k_d$ is:

$$k_d = \frac{(d+2)}{2} \left( \frac{2\gamma d}{d} \right)^{d/(d+2)} \quad (5.7)$$

where $\gamma_d = \frac{1}{2} \xi_{d-1}^2 \omega_d^{2/d}$ is the lowest eigenvalue of the operator $-\frac{1}{2} \Delta$ for the sphere of unit volume in $d$ dimensions with zero boundary values (absorbing boundary condition), with $\omega_d = \pi^{d/2}/\Gamma(1 + d/2)$ the volume of a $d$-dimensional sphere of unit radius, and $\xi_p$ the smallest zero of the Bessel function of the first kind of order $p$. The stretched exponential at large times is a consequence of diffusion through large trap free regions which appear with exponentially small probability. The correct asymptotic result can be obtained also through a simple heuristic argument (Grassberger and Procaccia, 1982). Since we are interested in long times and large trap free regions we can treat the problem in the continuum limit and replace a random walk with Brownian motion with diffusion constant $D = 1/(2d)$. The probability that there are no traps inside a sphere of radius $R$ is $(1 - c)^{\omega_d R^d} = \exp(-\lambda \omega_d R^d)$. The survival probability of a walker originally at the origin of the sphere satisfies the diffusion equation with absorbing boundary conditions. At long times the survival probability is dominated by the normal mode with the slowest decay $\exp(-n D \xi_{d-1}^2 / R^2)$. The joint probability for there to be a trap-free sphere of radius $R$, and for the diffusing walker to stay inside the sphere for at least $n$ steps is thus:

$$\Phi_{ON}(R, n) = \exp \left[ -\lambda \omega_d R^d - n D \xi_{d-1}^2 / R^2 \right] \quad (5.8)$$

The joint probability is maximized at the radius $R = (2n D \xi_{d-1}^2 / d \lambda \omega_d)^{1/(d+2)}$, which
results in

\[ \Phi_{ON}(n) \sim \exp \left[ -\frac{(d+2)}{2} \left( \frac{2}{d} n D \xi_{d}^{2} \right)^{\frac{d}{d+2}} \left( \lambda \omega_d \right)^{\frac{2}{d+2}} \right] . \] 

Note that this is identical to the Donsker-Varadhan result.

Interestingly, the survival probabilities for different concentrations are connected through a scaling function (Barkema et al., 2001; Gallos and Argyrakis, 2001b). The scaling ansatz for a function \( F_d \), which connects the small \( n \) (Rosenstock) and large \( n \) (Donsker-Vardahan) limits for all trap concentrations \( \lambda \) is:

\[ \ln \Phi_{ON} = -n^{\alpha_d} F_d \left( \lambda n^{\beta_d} \right) , \]  

with \( F_d(x) = a_d x \) for small \( x \) and \( F_d(x) = k_d x^{2/(d+2)} \) for large \( x \). Exponents \( \alpha_d = (d-2)/d \) and \( \beta_d = 2/d \) for \( d \geq 3 \) (\( \alpha_1 = 0, \beta_1 = 1/2 \)) are chosen such that the survival probability satisfies equations (5.2) and (5.6) for small and large \( n \). In two dimensions logarithmic factors must be added to the scaling function to account for logarithmic factors in the Rosenstock regime, resulting in

\[ \ln \Phi_{ON} = -\ln(8n) F_2 \left( \frac{\lambda n}{\ln^2(8n)} \right) . \] 

The transition from Rosenstock to Donsker-Vardahan regime occurs at \( a_d x_d^* \approx k_d x_d^{2/(d+2)} \), where \( x_d^* = \lambda n^{\beta_d} \). The values \( x_c^* \) for the first three dimensions are \( x_1^* = 27\pi^{-7/2}/(128\sqrt{2}) \approx 8.2 \), \( x_2^* = \xi_0^2/\pi \approx 1.84 \) and \( x_3^* = (5^{5/8} P(0; 1)^5/2^{13}3^{8})^{1/3} \approx 26.3 \), which are slightly bigger then the ones reported by numerical simulations (Barkema et al., 2001). Note that for \( d \geq 2 \), the survival probability at the transition depends on the trap concentration, and except for the densely trapped cases of \( c \approx 1 \), the survival probability at the transition is too small for the Donsker-Vardahan limit to be observed by direct numerical simulations of random walks (den Hollander and Weiss, 1994). Several numerical techniques have been developed that allow exploration of the Donsker-Vardahan limit, e.g. by exact enumeration of clusters of non-trapping sites (Havlin et al., 1984), using the complete distribution of \( S_n \) (Gallos and Argy-
Figure 5-1: Collapse of survival probabilities: Numerical simulations of the survival probability \( \Phi_{ON} \) for different trap concentrations for \( d = 1, 2, 3 \) obey the scaling shown in equations (5.10) and (5.11). Colored solid lines from left to right correspond to \( c = 10^{-5} \) (dark blue), \( 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 0.9, 0.99 \) (dark red). The Rosenstock approximation (dashed green line) works well at small \( \lambda \) and \( n \), while the Donsker-Varadhan limit (dashed black line) is reached only in one dimension. In two and three dimensions only the early stages of the transition to the long time limit are seen. For clarity, results for different dimensions \( d \) are shifted as indicated by the vertical axis label. This figure is adapted from Figure 1 in (Košmrlj, 2011).

Figure 5-1 summarizes all results for the survival probability from this section. We calculated survival probabilities at different trap concentrations ranging from \( c = 10^{-5} \) to \( c = 0.99 \) by directly simulating random walks and averaging over different trap configurations. Survival probability curves for different concentrations collapse on a single curve confirming the scalings from equations (5.10) and (5.11). The Rosenstock approximation works very well for small values of \( x \), except for very small \( n \), where the asymptotic limit of \( \langle S_n \rangle \) is not yet reached. The Donsker-Varadhan limit is only observed in one dimension, while in two and three dimensions only the early stages of the transition to the long time limit are seen.
Figure 5-2: Model. A random walker (black circle) steps on a lattice with randomly placed immobile traps. At each step a walker hops to a randomly chosen nearest neighbor site. If it steps on a trap in an absorbing “on” state (red), it dies, while if it attempts to step on a trap in a “reflective” state (blue) it remains at the initial site. Traps stay in the “on” ("off") state for typical time $n_{ON}$ ($n_{OFF}$) before switching to the other state. This figure is adapted from Figure 2 in (Kosmrlj, 2011).

5.2 Intermittent traps

In the previous section we discussed a classical trapping problem in which traps are always in the absorbing ("on") state. In this section the second reflective ("off") state is added to the model (Figure 5-2). If a random walker attempts to step on a trap in the “off” state it remains at the initial site, while it dies if it attempts to step on a trap in the “on” state. Switching times of trap states are exponentially distributed random variables, with traps staying in the “on” ("off") state for characteristic time $n_{ON}$ ($n_{OFF}$). The probability that a trap is initially in the state $x$ ("on", "off") is

$$p_x = \frac{n_x}{n_{ON} + n_{OFF}}.$$  \hspace{1cm} (5.12)

To numerically obtain survival probabilities, we directly simulate the process. We start with a large $d$-dimensional lattice grid (containing $\sim 10^{10}$ lattice points) that is divided into smaller hypercubes, each of which contains $\sim 1000$ lattice points. The grid is large enough that a random walker is always trapped before reaching the edges of the grid. To save computation time, a random configuration of traps inside a hypercube is generated only after a random walker first enters that hypercube. The
state of each trap is determined only after the random walker tries to jump on it and the state is chosen according to the probabilities in equation (5.12). After that the trap switches between the states with random exponentially distributed switching times. We simulate a random walk and measure how long the walker stays alive. This process is repeated many times with different trap configurations to obtain the survival probability \( \Phi(n, c, n_{\text{ON}}, n_{\text{OFF}}) \) after \( n \) steps with trap concentration \( c \).

What is the survival probability \( \Phi \) of a random walker in this case, and how is it related to \( \Phi_{\text{ON}} \) where traps are turned “on” all the time? The survival probability \( \Phi \) is bounded from below by \( \Phi_{\text{ON}} \), because having all traps “on” increases the probability of trapping

\[
\Phi \geq \Phi_{\text{ON}}. \tag{5.13}
\]

At large number of steps \( n \) in the Donsker-Varadhan limit, the survival probability is dominated by large trap free regions. The slow step is diffusion through this region, while finding the trap in an “on” state outside this region is quicker. Therefore at large \( n \)

\[
\Phi \simeq \Phi_{\text{ON}}. \tag{5.14}
\]

For small \( n \) and low trap concentrations \( c \) (\( c(S_n) \ll 1 \)), the probability of trapping \( (1 - \Phi) \) is dominated by the first collision with a trap, which must be turned on:

\[
1 - \Phi \simeq p_{\text{ON}} \left(1 - \Phi_{\text{ON}}\right). \tag{5.15}
\]

Figure 5-3 shows survival probabilities in \( d = 1, 2, 3 \) for different values of \( n_{\text{ON}} \) and \( n_{\text{OFF}} \). The trapping relation (Equation 5.15) works well at short times \( n \). Crossover to the long time Donsker-Varadhan limit (5.14) is seen only in one dimension and the transition is slower in higher dimensions. When switching times \( n_{\text{ON}} \) and \( n_{\text{OFF}} \) are slow we notice that the survival probability plateaus at some level, before it starts dropping again. The reason for this is a caging effect described below.

In \( d = 1 \) a random walker is always bounded by the nearest left and right traps. When switching between trap states is very slow, the trap configuration is “frozen”
Figure 5-3: Survival probability $\Phi$ in $d = 1, 2, 3$ ($c = 0.1, n_{\text{OFF}}/n_{\text{ON}} = 1$). Colored solid lines correspond to different values of $n_{\text{OFF}}$; the darkest red color on the right corresponds to the largest $n_{\text{OFF}} = 10^7$ and neighboring curves differ by factors of 10 in $n_{\text{OFF}}$. For small $n$, the trapping relation (Equation 5.15, dotted green line) works well. $\Phi_{\text{ON}}$ (solid black line) represents a lower bound for $\Phi$, which is reached in the Donsker-Varadhan limit of large $n$. Crossover to the Donsker-Varadhan limit is slower in higher dimensions. When switching times $n_{\text{ON}}$ and $n_{\text{OFF}}$ are slow (yellow, orange, red) the survival probability plateaus before it starts dropping again. This figure is adapted from Figure 3 in (Košmrlj, 2011).

Figure 5-4a: Survival probability $\Phi$ in $d = 1, 2, 3$ ($c = 0.1, n_{\text{OFF}}/n_{\text{ON}} = 1$). Colored solid lines correspond to different values of $n_{\text{OFF}}$; the darkest red color on the right corresponds to the largest $n_{\text{OFF}} = 10^7$ and neighboring curves differ by factors of 10 in $n_{\text{OFF}}$. For small $n$, the trapping relation (Equation 5.15, dotted green line) works well. $\Phi_{\text{ON}}$ (solid black line) represents a lower bound for $\Phi$, which is reached in the Donsker-Varadhan limit of large $n$. Crossover to the Donsker-Varadhan limit is slower in higher dimensions. When switching times $n_{\text{ON}}$ and $n_{\text{OFF}}$ are slow (yellow, orange, red) the survival probability plateaus before it starts dropping again. This figure is adapted from Figure 3 in (Košmrlj, 2011).

for $n \ll n_{\text{ON}}, n_{\text{OFF}}$. If at least one of the bounding traps is in “on” state, a random walker dies at that trap. With probability $p_{\text{OFF}}^2$ both bounding traps are in the “off” state, forming a cage. The walker diffuses inside this cage until one of the traps turns “on” and then it is quickly absorbed. The “survival” probability of the cage is:

$$\Phi_c(n) = p_{\text{OFF}}^2 \exp[-2n/n_{\text{OFF}}], \quad (5.16)$$

where the exponential factor represent the probability that both bounding traps are still in the “off” state after $n$ steps. After the survival probability $\Phi$ of a walker drops to the probability of caging, it is well approximated by the above equation (Figure 5-4a). $\Phi_c$ decays faster than the stretched exponential in the Donsker-Varadhan limit, which comes from large trap free regions. Therefore at large $n$ the survival probability is still dominated by the large trap free regions (Figure 5-3).

In $d > 1$ traps always form a cage when the trap concentration $c$ is above the percolation threshold. However, even at low trap concentration $c$, with low probability
there are some trap configurations that form small cages around the origin. If at least one of the traps forming a cage is in the "on" state, then a random walker dies quickly. But if all cage forming traps are in the "off" state, and the typical switching time $n_{OFF}$ is slow, then a random walker diffuses inside the cage until one of the traps switches to the "on" state. The average survival probability of a random walker that is caught in a cage is thus

$$\Phi_c = \sum_{k, \ell} (A_{\ell,k}(1-c)^k c^\ell) \times (p_{OFF}^\ell \exp[-\ell n/n_{OFF}]),$$  \hspace{1cm} (5.17)

where $\ell$ is the number of traps forming a cage, and $k$ the number of empty sites inside the cage excepting the origin (because the origin - the initial location of the random

Figure 5-4: Caging effect: a) Survival probabilities in $d = 1$ ($c = 0.1, n_{OFF}/n_{ON} = 1$) agree well with the probability of caging (Equation 5.16, dashed yellow line) when switching times are slow (colored solid lines in a) and c) are the same as in Figure 5-3). The solid yellow line is initially close to the probability of caging, but at large $n$ it starts deviating and approaches the stretched exponential of the Donsker-Varadhan limit. b) Configuration of cages in $d = 2$ with the smallest number $\ell$ of cage forming traps. A random walker (black) sits at the origin and is surrounded by traps in the “off” state (blue). $A_{\ell,k}$ gives the number of equivalent cage configurations. c) Survival probabilities in $d = 2$ ($c = 0.8, n_{OFF}/n_{ON} = 10$) agree well with the probability of caging (Equation 5.17) when switching times are slow. For large $n$ it is sufficient to keep only one (dashed red line) or two (dashed black line) terms in Equation (5.17). Because $c = 0.8$ and $p_{OFF} = 10/11$ are close to 1, more terms need to be kept for accurate description of the caging probability at small $n$ (dotted yellow line, up to $\ell = 11$). This figure is adapted from Figure 4 in (Kosmrlj, 2011).
walker - is always assumed to be empty). The first factor represents the probability of cage formation with $A_{\ell,k}$ being the combinatorial number of configurations with the same number of empty sites and cage forming traps. The second factor is the probability that all cage forming traps are initially in the “off” state and remain in the same state for $n$ steps. In $d = 2$, the first few non-zero coefficients $A_{\ell,k}$ are $A_{4,0} = 1$, $A_{6,1} = 4$, $A_{7,2} = 12$ (Figure 5-4b) and in general for $d > 1$ the first two non-zero coefficients are $A_{2d,0} = 1$ and $A_{4d-2,1} = 2d$. When $c$ or $p_{OFF}$ are small, or $n \gg n_{OFF}$, it is enough to keep only the first few terms in equation (5.17), but if $c$ and $p_{OFF}$ are close to 1 and $n \lesssim n_{OFF}$ then we must keep more terms (Figure 5-4c). Like in one dimension, the survival probability $\Phi$ is well approximated for large $n_{OFF}$ by equation (5.17) after $\Phi$ drops to the probability of caging $\Phi_c$. $\Phi_c$ decays faster than the stretched exponential decay in the Donsker-Varadhan limit; therefore at large $n$ the survival probability should still approach $\Phi_{ON}$ (Figure 5-3).

5.3 Escape probability of autoimmune T cells

In conclusion, let us return to the question of escape probability of potentially autoimmune T cells from the thymus. First we estimate the relevant parameters in our model. The density of APCs in the thymus is 30,000–50,000 cells/mm$^3$ (Borgne et al., 2009). If we choose the lattice constant $a_0$ to be equal to the typical diameter of the antigen presenting cell $\sim 10\mu$m (Borgne et al., 2009), the trap concentration is estimated at $c \sim 0.03$–0.05. From the T cell diffusion constant, $D \sim 100\mu m^2 min^{-1}$ (Borgne et al., 2009), we estimate the hopping time of a random walker as $t_1 = a_0^2/(6D) \sim 0.2$min. T cells diffuse around the thymus for around 4–5 days (Klein et al., 2009), which is equivalent to $n \sim 30,000$–40,000 steps of a random walker. A typical lifetime of a self-peptide on the surface of APCs is on the order of hours (Henrickson et al., 2008), which corresponds to $n_{ON} \sim 100$. Since many self-peptides are simultaneously presented on the surface of APCs, the effective $n_{ON}$ could be even longer. Different T cells are represented with different $n_{OFF}$. A potentially autoimmune T cell is more likely to bind strongly to APCs and is thus associated with lower values of
Let us first estimate the escape probability for an extremely autoimmune T cell, which reacts to every host cell. Such a T cell is characterized by \( n_{\text{OFF}} = 0 \) and \( \Phi = \Phi_{\text{ON}} \). Since \( x = \lambda r^{2/3} \sim 30-60 \) is larger than the location \( x_3^* = 26 \) of the transition to the Donsker-Varadhan limit, we use the Donsker-Varadhan result to estimate \( \Phi_{\text{ON}} \sim 10^{-400} - 10^{-300} \). This implies that an extremely autoimmune T cell cannot escape from the thymus.

In reality autoimmune T cells do not react to every host cell, but are directed toward particular tissues. Since tissue specific peptides are presented on 1-3% of APCs in the thymus (Klein et al., 2009), such autoimmune T cells are characterized by \( n_{\text{OFF}} \sim 3,000-10,000 \) in our model (\( \rho_{\text{ON}} \sim 1-3\% \)). The model estimates that the escape probability of T cells specific for tissues that are represented with fewer peptides (\( \rho_{\text{ON}} \sim 1\% \)) is \( \sim 10^{-6} - 10^{-3} \) and the escape probability of T cells specific for tissues that are represented with more peptides (\( \rho_{\text{ON}} \sim 3\% \)) is \( \sim 10^{-20} - 10^{-10} \). The estimates are very crude, because the survival probability depends exponentially on the parameters which are just approximately determined. However, we still see that the escape from the thymus is possible when tissue specific peptides are not presented in big amounts. This may be why the immune system contains specialized regulatory T cells, which try to shut down the response from autoimmune T cells.

Finally, we note that the caging effect is not relevant for thymic selection: since the trap concentration \( c \) is low, even small cages appear with very small probability and thus the caging effect is not important on the relevant time scales. In any case, T cells can never be trapped inside cages, as they are delivered to the thymus via blood vessels and small cages around the entrance would completely block the flow of T cells through the thymus.

In this paper we introduce to study properties of a model initially motivated by T cell survival. While the main body of the text explores the full parameter range of this model, some of the results do not apply to the relevant parameters for the thymic selection. Furthermore for thymic selection the model parameters are only approximately known. Because the escape probability of autoimmune T cells depends exponentially on these parameters, we can only determine the order of magnitude for
the escape probability, which could probably be estimated also with simpler models.
Chapter 6

Conclusions

In this thesis, I presented studies that enable an understanding of how the developmental processes in the thymus shape the T cell repertoire such that adaptive immunity exhibits both remarkable pathogen specificity and the ability to combat myriad pathogens. We proposed a statistical mechanical model of thymic selection which suggests that the peptide binding residues of selected T cell receptors are enriched with weakly interacting amino acids; a result that is in agreement with available receptor structures (Chapters 2 and 3 and (Košmrlj et al., 2008, 2009)). Sufficiently strong binding for T cell receptor recognition of a foreign peptide is thus achieved via many moderately strong bonds, each of which makes a significant contribution to binding, again consistent with experiments (Huseby et al., 2006). Breaking any of these bonds by mutating the foreign peptide may reduce the binding strength and prevent recognition from the same T cell receptor, because the binding strength falls below the sharp threshold for recognition. In this sense, T cell receptors are specific for the foreign peptide and do not recognize the peptide after single amino acid mutations. Once T cell receptors recognize foreign peptides in this statistical fashion, then the interaction can persist to allow enough time for possible modest induced fit changes in T cell receptor conformation to stabilize the interaction even further (Gakamsky et al., 2007; Wu et al., 2002).

The thymic selection model has some implications also for human diseases. Most people express different sets of MHC types, and people expressing a particular MHC
type are genetically predisposed for certain diseases. In particular people expressing the HLA-B57 type of MHC molecule are more likely to control the HIV infection even without therapy, but they are also more prone to autoimmunity. Undoubtedly there are many reasons for this tendency, but we explored only the effect of thymic selection (Chapter 4). Our bioinformatics study showed that the HLA-B57 molecule binds fewer types of self peptides compared to other MHC molecules. In agreement with experiments, our thymic selection model suggests that in these people T cell repertoires are more cross-reactive, and are more effective at recognizing the mutating viral peptides. At the same time, more cross-reactive T cells are also more likely to trigger against self peptides, because the thymic selection is not perfect. We predicted that MHC molecules that bind fewer (more) peptides are better (worse) for the control of HIV. Our collaborators tested this prediction on a large cohort of HIV infected people and our prediction seems to be valid for the MHC types that bind very few or bind a lot of self peptides (Chapter 4 and (Kosmirj et al., 2010)), while for people with other MHC types other effects might be more relevant to HIV control. Thus we have identified one new effect in people expressing HLA-B57 that could help them control the HIV infection; undoubtedly other effects also contribute to these individuals’ ability to control HIV.

In the future it would be important to test the combined effect of MHC types (each human inherits one HLA-B type from mother and one from father) on the control of HIV. Even though T cell receptors interact productively with only one MHC type, the dynamics of T cells are coupled through an MHC competition for peptides and a competition for the presentation of peptide-MHC complexes on the surface of infected cells. The combined effect of MHC types could be studied, by adapting the host–pathogen interaction model presented in this thesis. The literature (Carrington et al., 1999) report that the homozygosity of MHC types (inherited HLA-B types are the same) speeds up the progression to AIDS. The progression to AIDS is especially fast when a human inherits non-protective HLA-B35 type from both parents. It would be interesting to study what are the effects of combinations of protective/protective, non-protective/non-protective, and especially protective/non-protective MHC types
on the control of HIV.

Common vaccination strategies are infection with a live crippled virus or usage of mosaic vaccines (cocktail of viral peptides). Immune system clears the vaccine and generates a pool of memory T cells, which are specific for viral/vaccine peptides. Upon viral infection these memory T cells generate faster and stronger immunological response. By combining the T cell development model and the model of host-pathogen interaction, we could study the efficiency of different mosaic vaccines and identify the optimal combination of viral peptides used in vaccines.

T cells that could potentially target healthy host cells are eliminated during the thymic selection. However, the flexible adaptive immune system can also go awry, and many diseases (e.g., multiple sclerosis and type I diabetes) are the consequence of the T cells failing to discriminate between markers of self and non-self. The suffering caused by autoimmune diseases and the need to combat diverse infectious agents have motivated a great deal of experimental research aimed toward understanding how the adaptive immune system is regulated, and indeed, some spectacular discoveries have been made. Yet, an understanding of the mechanistic principles that underlie an autoimmune response remain elusive. This is especially true with regard to the origins and triggers of autoimmune diseases. Some insights could be obtained by using statistical mechanics tools, similar to the ones developed in this thesis. For example:

- By adapting the thymic selection model, to include the variability in expression levels of different types of self peptides in the thymus, one could get insight into questions, such as: What leads to a higher frequency of escape of autoreactive cells which target cells of the nervous system and the pancreas in the case of autoimmune diseases like multiple sclerosis and type I diabetes? Why are the escaping autoreactive cells attacking only particular tissues?

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

- Why are frequencies of expressed HLA molecules different among different geographical populations? What gives certain HLA molecules a selection advantage in populations with different environments and history? This problem could be studied by adapting the stepping-stone model of spatio-temporal evolution in population genetics (Kimura and Weiss, 1964).

Many problems in immunology at the molecular, cellular, and organism level remain unsolved and present future research opportunities. However, a key challenge just beginning to be addressed is the following. Much progress (theoretical and experimental) has been made to describe host-pathogen dynamics in humans (especially in the context of HIV). Basic molecular and cellular immunology has focused largely on experimental models that are inbred mice or cell lines. This is because these models can be readily manipulated to test ideas (such as predictions emerging from statistical mechanical theory and computation). A fundamental challenge is to bridge the gap between basic molecular discoveries and the immune response of individual humans with specific genes (e.g., people have diverse MHC genes). Much is to be gained by making the two meet, and understanding the effects of fluctuations in host genetics is an interesting statistical mechanics problem.
Appendix A

Robustness of the thymic selection model

In this appendix we demonstrate a robustness of the results of the thymic selection model presented in Chapter 2 to variations of model parameters, such as the interaction potential $J$ (Figures A-1–A-4), the softness of positive and negative selection thresholds (Figure A-5), amino acid frequencies of self peptides and TCRs (derived from human vs mouse proteome, Figure A-6), TCR binding strengths to MHC (Figures A-7). Additionally we show how the results depend on the number of self peptides in the thymus (Figure A-8) and how the selection probability of TCRs depend on its binding strength to MHC (Figure A-9). We also show that the results of the analysis of TCR-pMHC crystal structures are robust to the definition of contact between amino acids on the peptide and TCR (Figure A-10).

A.1 Robustness to changes in interaction potential.

In all calculations reported in Chapter 2, the MJ matrix (Miyazawa and Jernigan, 1996) was used to determine the interaction free energy between peptide contact residues of the TCR and peptide amino acids. Here, we explore what happens if we
Figure A-1: Amino acid frequencies in selected TCRs for the MJ matrix $J$. (a) Color representation of the MJ interaction matrix in units of $k_B T$. (b) Amino acid frequencies in selected TCRs for selection against different numbers of self peptides $M$. The abscissa is a list of amino acids ordered according to the maximum free energy with which it interacts with all other amino acids.

Use semi-random symmetric matrices with the same values of mean and variance as the MJ matrix $J$ and controlled differences between the largest values in each row. In order to do that we construct a symmetric random matrix $R$, by first assigning each matrix element a random uniformly distributed value from -1 to 0 and then rescaling the matrix, such that the mean and the variance of the elements of matrices $J$ and $R$ are the same. To control the differences between amino acid strengths we use matrices of the form $(1 - x)J + xR$, where $x \in [0, 1]$.

As shown in Figure A-1a there is a clear gradation of interaction free energies (color scale in $k_B T$ units) in the MJ matrix, from the strong (lower left, blue color) to weak (upper right, red color), enabling a clear ordering of the amino acids. For the MJ matrix, the order of amino acids obtained by using the average interaction free energy with other amino acids or that obtained by using the strongest interaction with other amino acids is quite similar. Therefore, the computational results are unchanged from that shown in Figure 2-5a if results using the MJ matrix are graphed with the amino acids ordered according to their average interaction with other amino acids. For random matrices (e.g., Figures A-2 and A-3), the average value of an amino acids
interaction energies with other amino acids and the strongest interaction of this amino acid with all others are not correlated. Our analytical calculation (see Chapter 3) shows that ordering amino acids according to their strongest interaction with other amino acids is appropriate when there are many types of peptides in the thymus. Therefore, we use this criterion in Figure A-1b (MJ matrix) and Figures A-2b and A-3b (different random potentials). Results for the random potentials are qualitatively similar to that for the MJ matrix when this criterion is used. We varied the random potential by varying the difference between the maximum interaction free energies characterizing the strongest and weakest amino acids (L and K), which is around $4k_B T$ for the MJ matrix. When we make this difference smaller (e.g., $2k_B T$ as in Figure A-2a), there is no clear trend of amino acid composition when TCR develop in a thymus with a small number of types of peptides (Figure A-2b). Importantly, for many types of peptides in the thymus, the qualitative trends obtained for the MJ matrix are recovered. This is also true for even smaller differences between the strongest and weakest amino acids (e.g., $0.6k_B T$) in Figure A-3. For random potentials, there are more "bumps" in the distribution, but these disappear if an even larger number of
endogenous peptides are displayed in the thymus. For non symmetric interaction matrices, statistical properties of selected TCRs are also similar to that we have reported, and the order of amino acids is determined by the strongest interactions with other amino acids (data not shown).

More complex interactions between peptide contact residues of TCRs and peptide amino acids are used to check the robustness of our results. For example we present the results (Figure A-4) for the following more complex potential, which includes interactions with “nearest neighbor” amino acids

\[
E_{\text{int}}(\vec{t}, \vec{s}) = E_c + \sum_{i=1}^{N} \left[ J(t_i, s_i) + \frac{1}{2} \{ J(t_i, s_{i+1}) + J(t_i, s_{i-1}) \} \right].
\]  

(A.1)

\(J(t_i, s_i)\) is the interaction free energy between the \(i\)th amino acids on the variable part of the TCR \((t_i)\) and the peptide \((s_i)\), respectively, and \(N\) is the length of the variable regions. In fact, the statistical properties of the TCR repertoire (Figure A-4)
Figure A-4: The qualitative features of the post-thymic selection TCR repertoire are robust to more complex interactions (Equation A.1) between peptide contact residues of the TCR and peptide amino acids. We show: (a) the number of important contacts, (b) the amino acid composition of selected TCRs, (c) the distribution of contact free energies between selected TCRs and antigenic pMHC and (d) the amino acid composition of recognized antigenic peptides ($E_n - E_c = -75k_BT, E_n - E_p = -5k_BT$).

remain unchanged for any bilinear combination ($C_{\alpha\beta} \geq 0$)

$$E_{\text{int}}(\vec{r}, \vec{s}) = E_c + \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} C_{\alpha\beta} J(t_\alpha, s_\beta).$$  

(A.2)
A.2 Robustness to changes in other parameters.

In this section we demonstrate a robustness of the results of the thymic selection model presented in Chapter 2 to variations of model parameters, such as the softness of positive and negative selection thresholds (Figure A-5), amino acid frequencies of self peptides and TCRs (derived from human vs mouse proteome, Figure A-6), TCR binding strengths to MHC (Figures A-7). Additionally we show how the results depend on the number of self peptides in the thymus (Figure A-8) and how the selection probability of TCRs depend on its binding strength to MHC (Figure A-9). We also show that the results of the analysis of TCR-pMHC crystal structures are robust to the definition of contact between amino acids on the peptide and TCR (Figure A-10).
Figure A-5: Soft thresholds for positive and negative selection. The qualitative properties of the post-thymic selection TCR repertoire are robust to the nature of thresholds for positive and negative selection. We show: (a) the number of important contacts, (b) the amino acid composition of selected TCRs, (c) the distribution of contact free energies between selected TCRs and antigenic pMHC with a soft threshold for positive selection and (d) the amino acid composition of recognized antigenic peptides. The statistical properties of the TCR repertoire remain unchanged upon introduction of soft thresholds for positive and negative selection. The statistical properties of the TCR repertoire remain unchanged upon introduction of soft thresholds for positive and negative selection. The statistical properties of the TCR repertoire remain unchanged upon introduction of soft thresholds for positive and negative selection. The statistical properties of the TCR repertoire remain unchanged upon introduction of soft thresholds for positive and negative selection (\(E_n - E_c = -40k_BT\), \(E_n - E_p = -5k_BT\), \(\sigma_n = 0.2k_BT\), \(\sigma_p = 2k_BT\); the threshold softness parameters \(\sigma_n\) and \(\sigma_p\) are defined in Section 2.6).
Figure A-6: Thymic selection using amino acid frequencies from mouse proteome. Distributions of (a) the number of important contacts, (b) amino acid composition of selected TCRs, (c) contact free energies between selected TCRs and antigenic pMHC, and (d) amino acid composition of recognized antigenic peptides are similar whether using amino acid frequencies from mouse or human proteome to generate TCRs and self peptides. ($E_n - E_c = -40k_BT$, $E_n - E_p = -5k_BT$).
Figure A-7: Thymic selection results for different TCR binding strengths to MHC. We show: (a) the number of important contacts, (b) the amino acid composition of selected TCRs, (c) the distribution of contact free energies between selected TCRs and antigenic pMHC and (d) the amino acid composition of selected TCRs. For moderate changes in TCR binding strengths to MHC ($\Delta E_c > 0$ corresponds to stronger binding), the amino acid composition of selected TCRs (b) changes slightly, such that the properties of TCR recognition of antigenic peptides (a, d) remain unchanged. ($E_n - E_c = -40k_BT$, $E_n - E_p = -5k_BT$).
Figure A-8: Results of thymic selection as a function of the number of self peptides $M$ in the thymus. (a) The number of important contacts increase with number of self peptides presented in the thymus. (b) The dependence of the amino acid distribution of selected TCR sequences as a function of the number of self peptides in the thymus. (c) The distribution of contact free energies between selected reactive TCRs and antigenic peptides. (d) The dependence of the amino acid distribution of recognized antigenic peptides as a function of the number of self peptides in the thymus. Increasing the number of self peptides in the thymus results in more moderate contacts and less weak and strong contacts. These results (particularly b) show that as long as there are $\geq 100$ types of endogenous pMHC in the thymus, the qualitative results reported in Chapter 2 would be obtained. ($E_n - E_c = -40k_B T$, $E_n - E_p = -5k_B T$).
Figure A-9: Probabilities of TCR selection. (a) Fraction of selected TCRs ($P_{\text{sel}}(E_c)$) against one self-peptide (black curve) and many types of self-peptides (blue curve, $M = 10,000$) as a function of the threshold for negative selection $E_n - E_c$, whereas the gap between thresholds for negative and positive selection is kept constant at $E_n - E_p = -5k_BT$. At small values of $E_n - E_c$ (strong binding of TCR to MHC) negative selection is dominant – dotted lines show fraction of TCRs that are not negatively selected. At large values of $E_n - E_c$ (weak binding of TCR to MHC) positive selection is dominant – dashed lines show fraction of TCRs that are positively selected.

(b) Numerically obtained probability density function of TCR binding strengths to MHC ($E_c$ values) in selected TCRs ($\rho_{\text{sel}}(E_c)$) for selection against $M$ self peptides. Before selection $E_c$ values are taken to be uniformly distributed ($\rho_0(E_c)$) on a narrow interval $(E_{c,\text{min}}, E_{c,\text{max}}) = (-35.5, -30.5)k_BT$. The probability density function of TCR binding strengths to MHC after selection can be obtained analytically $\rho_{\text{sel}}(E_c) = \rho_0(E_c)P_{\text{sel}}(E_c)/\int \rho_0(E)P_{\text{sel}}(E)dE$ (cf. the left panel (a)). TCRs that bind more strongly to MHC may also be selected, when selected against fewer types of self peptides.
Figure A-10: Frequency distribution of amino acids in TCR that are in contact with peptide. The ratio of amino acid frequencies derived from the list of amino acids of TCRs in contact with peptides calculated from 18 available crystal structures of TCR–pMHC(I) complexes with respect to the amino acid frequencies from human proteome are presented in these graphs. The residues are said to be in contact with each other if the $C_{\alpha}–C_{\alpha}$ distance is < 6.5 Å (black points). In a separate analysis, any two residues are defined to be in contact if a water molecule cannot fit between them (blue points). The dominance of weakly interacting amino acids is robust to the definition of contact between residues. (a) The abscissa is divided into the two types of strong amino acids (IVYWREL) and weak amino acids (QSNTAG) according to reference (Zeldovich et al., 2007) (b) The amino acids on abscissa are ordered from strongest (L) to weakest (K) according to the strongest interaction with another amino acid using the MJ matrix. This ordering presents charged amino acids (REDK) as weak. In contrast, according to reference (Zeldovich et al., 2007), amino acids R and E are strong, and amino acids D and K are not weak.
Appendix B

Accuracy of prediction algorithms for peptide binding affinities to HLA and Mamu alleles

There are at present several HLA–peptide binding prediction methods. The performance of these algorithms to identify new epitopes has recently been benchmarked against experimental data (Peters et al., 2006). In general, artificial neural networks, ANN, (Gulukota et al., 1997) and the stabilized matrix method, SMM, (Peters et al., 2003) were found to be superior to other methods (Peters et al., 2006). We used the ANN and the SMM (versions 2009-09-01 and 2007-12-27) prediction tools provided by the IEDB (Peters et al., 2005). Accuracy of prediction tools was tested against experimental data downloaded from the IEDB in September 2009. These experimental data were obtained by two methods: competition assays, in which purified MHC and radioactive labelling are used; and association studies, in which purified MHC and fluorescence labelling are used. Data obtained from the two methods show significant correlations of measured binding affinities (as measured by half-maximum inhibitory concentration (IC$_{50}$) and half-maximum effective concentration (EC$_{50}$)) (Rao et al., 2009). Prediction tools were tested against experimental data for accuracy of classifying peptides into binders (IC$_{50} < 500$ nM) and non-binders (IC$_{50} \geq 500$ nM); the chosen thresholds are commonly accepted values (Rao et al.,
We also tested how well these tools predict absolute measured affinity values, not just classification of binders and non-binders, which is dependent on the chosen thresholds. The accuracy of the prediction tools thus determined are summarized in Table B.1 and Figures B-1-B-3. We excluded all HLA and Mamu alleles for which there was not enough experimental data (at least 50 binders and 50 non-binders) or prediction tools were not sufficiently accurate. For each HLA and Mamu allele, the most accurate prediction tool was used to predict the fraction of unique peptides derived from the human and macaque proteome (Homo_sapiens.GRCh37.55.pep.all.fa and Macaca_mulatta.MMUL 1.56.pep.all.fa obtained from Ensembl (Hubbard et al., 2009)) that can bind to that allele. We focused only on the binding abilities of peptides of 9 amino acids to HLA molecules, because there is not enough experimental data available for the binding affinities of peptides of 8, 10 and 11 amino acids to HLA-B*5701 and the other relevant HLA-B alleles that emerged from our analyses (HLA-B*2705, HLA-B*0702 and HLA-B*3501).

B.1 Measures of accuracy of prediction algorithms

For each HLA and Mamu allele we have analyzed the accuracy of four predictive algorithms available from the Immune Epitope Database (IEDB): ANN and SMM (versions 2009-09-01 and 2007-12-27). Accuracy was tested against experimental data (downloaded from the IEDB) of measured peptide binding affinities (IC$_{50}$) to HLA and Mamu molecules. Only those HLA and Mamu alleles were kept in the analysis for which there was enough experimental data (at least 50 binders and 50 non-binders).

First we tested how good the predictive algorithms are in classifying peptides into binders (IC$_{50}$ < 500 nM) and non-binders (IC$_{50}$ ≥ 500 nM). We counted the number of true positives $T_p$ (correctly predicted binders), true negatives $T_n$ (correctly predicted non-binders), false positives $F_p$ (incorrectly predicted binders) and false negatives $F_n$ (incorrectly predicted non-binders). The accuracy of the algorithm is defined as

$$ACC = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}. \quad (B.1)$$
Most algorithms for which there were sufficient experimental data were very accurate (more than 80%, Table B.1). Commonly used measure of accuracy is also Matthews correlation coefficient:

\[ MCC = \frac{T_p T_n - F_p F_n}{\sqrt{P M P' M'}} \]  

where \( P (M) \) and \( P (M) \) are numbers of experimental binders (non-binders) and predicted binders (non-binders) respectively. The closer the \( MCC \) is to the value 1, the higher the accuracy of the prediction algorithm. Most algorithms had a \( MCC \) value in the range of 0.6 to 0.9 (Table B.1). The accuracy of predictive algorithms is often displayed with a Receiver operating characteristic (ROC) curve (Swets, 1988). The ROC curve shows how the true positive rate, \( TP/(TP + FN) \), and the false positive rate, \( FP/(FP + TN) \), change as the threshold that separates predicted binders from non-binders is varied (see Figures B-1-B-3). A quantitative measure of the predictive algorithm accuracy is the area under the ROC curve (\( AUC \)). The closer the \( AUC \) is to the value 1, the higher the accuracy of the prediction algorithm. Most algorithms were very accurate and had an \( AUC \) value above 0.9 (Table B.1).

Second we tested how good the predictive algorithms are at determining the actual value of the binding affinity, \( IC_{50} \). Because binding affinities span a huge range of values, a commonly used difference of logarithms

\[ f(i) = \ln(\text{predicted } IC_{50}) - \ln(\text{measured } IC_{50}), \]  

was taken as a measure of the accuracy of predicted binding affinity of the \( i \)-th peptide. When experiments reported that binding affinity has a value greater than some value, \( LIC_{50} \), we defined

\[ f(i) = \begin{cases} 
0; & \text{predicted } IC_{50} \geq LIC_{50} \\
\ln(\text{predicted } IC_{50}) - \ln(\text{LIC}_{50}); & \text{predicted } IC_{50} < LIC_{50}
\end{cases} \]  

Similarly, when experiments reported that binding affinity has value lower than some
value, HIC50, the accuracy was defined as
\[
f(i) = \begin{cases} 
0; & \text{predicted IC50} \leq \text{HIC50} \\
\ln(\text{predicted IC50}) - \ln(\text{HIC50}); & \text{predicted IC50} > \text{HIC50}
\end{cases}
\] (B.5)

Overall accuracy of the predictive algorithm was determined from average bias \(\Delta \ln(y)\) and average root mean square error \(\sigma\):
\[
\Delta \ln(y) = \frac{1}{N} \sum_{i=1}^{N} f(i), \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} [f(i)]^2.
\] (B.6)

Table B.1 reports accuracies of all 4 predictive algorithms for each HLA-B allele and Mamu allele for which there was enough experimental data available. HLA-A alleles were not studied, because they are not associated with control of HIV (Price et al., 2009). In assessing the accuracy of the algorithms using Equations (B.1)-(B.6), we did not include experimental data for which even a bound (less than or greater than) for IC50 was not reported.

If average bias \(\Delta \ln(y) > 0 (\Delta \ln(y) < 0)\), then the predictive algorithm on average overestimates (underestimates) the value of binding affinity. Average bias of predictive algorithms can affect values of predicted fraction of peptides that can bind to a certain allele. In Table B.1, alleles for which there are no accurate predictive algorithms (all 4 predictive algorithms have large normalized bias \(|\Delta \ln(y)/\sigma| > 0.06\) are marked with white. The most accurate algorithm (bold font in Table B.1) for those alleles is taken to be the one with the least value of normalized bias \(|\Delta \ln(y)/\sigma|\). In Table B.1, alleles for which there is at least one accurate predictive algorithm (normalized bias \(|\Delta \ln(y)/\sigma| < 0.06\) are marked with yellow. If there is more than one accurate predictive algorithm for a given allele, the most accurate predictive algorithm (bold font in Table B.1) was selected to be the one with the least value of average root mean square error, \(\sigma\), among the accurate predictive algorithms (for all of which \(|\Delta \ln(y)/\sigma|\) was very small). Different choices of threshold for normalized bias that separate accurate and inaccurate predictive algorithms lead to only small changes in Table B.1. For example, if we choose the threshold to be 0.08, we get one more allele...
(HLA-B*1517) with at least one accurate predictive algorithm and the most accurate algorithm would change for two alleles (HLA-B*1517 and HLA-B*4403).

For each HLA and Mamu allele we used the most accurate predictive algorithm to predict the fraction of peptides derived from human and macaque proteome that can bind to given allele. There were \( \sim 10^7 \) (\( \sim 10^6 \)) unique peptide sequences in human (macaque) proteome. We used only HLA-B alleles (marked with yellow in Table B.1) for which there is at least one accurate predictive algorithm available to determine the typical binding fraction for HLA-B alleles (median – 0.013 and average – 0.015).
<table>
<thead>
<tr>
<th>allele</th>
<th>BF predictive algorithm</th>
<th>N</th>
<th>ACC</th>
<th>MCC</th>
<th>AUC</th>
<th>Δ ln(g)</th>
<th>σ</th>
<th>Δ ln(g)/σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-B*0702</td>
<td>smm (2009-09-01)</td>
<td>2301</td>
<td>94.1</td>
<td>0.840</td>
<td>0.982</td>
<td>0.108</td>
<td>1.065</td>
<td>0.101</td>
</tr>
<tr>
<td>HLA-B*0801</td>
<td>ann (2007-12-27)</td>
<td>1560</td>
<td>91.2</td>
<td>0.759</td>
<td>0.960</td>
<td>0.246</td>
<td>1.553</td>
<td>0.159</td>
</tr>
<tr>
<td>HLA-B*1501</td>
<td>ann (2007-12-27)</td>
<td>2342</td>
<td>87.3</td>
<td>0.713</td>
<td>0.946</td>
<td>-0.061</td>
<td>1.470</td>
<td>0.041</td>
</tr>
<tr>
<td>HLA-B*1503</td>
<td>smm (2009-09-01)</td>
<td>390</td>
<td>88.2</td>
<td>0.724</td>
<td>0.951</td>
<td>-0.191</td>
<td>1.700</td>
<td>0.112</td>
</tr>
</tbody>
</table>

**Table B.1: Accuracy of predictive algorithms for 9-mer peptide binding by HLA-B and Mamu alleles.**

### Notes:
- **BF**: fraction of 9-mer peptides (derived from human and macaque proteome) that are predicted to bind to HLA-B and Mamu alleles.
- **N**: number of available experimental measurements used to test prediction algorithms.
- **ACC**: % accuracy of classifying peptides into binders and non-binders (Equation B.1).
- **MCC**: Matthews correlation coefficient (Equation B.2).
- **AUC**: area under the ROC curve.
- **Δ ln(g)**: average bias and error of predicted binding affinity value IC50 (Equation B.6).
Table B.1 continued:

<table>
<thead>
<tr>
<th>Allele</th>
<th>BF</th>
<th>Predictive algorithm</th>
<th>N</th>
<th>ACC</th>
<th>MCC</th>
<th>AUC</th>
<th>Δ ln(y)</th>
<th>σ</th>
<th>(Δ \ln(y)/σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-B*5801</td>
<td>0.016</td>
<td>ann (2009-09-01)</td>
<td>1947</td>
<td>95.3</td>
<td>0.838</td>
<td>0.883</td>
<td>0.052</td>
<td>0.956</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>0.012</td>
<td>ann (2007-12-27)</td>
<td></td>
<td>95.0</td>
<td>0.830</td>
<td>0.880</td>
<td>0.016</td>
<td>1.016</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>0.017</td>
<td>smm (2008-09-01)</td>
<td></td>
<td>94.8</td>
<td>0.792</td>
<td>0.797</td>
<td>-0.039</td>
<td>1.194</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>0.014</td>
<td>smm (2007-12-27)</td>
<td></td>
<td>93.9</td>
<td>0.788</td>
<td>0.777</td>
<td>0.093</td>
<td>1.205</td>
<td>0.077</td>
</tr>
<tr>
<td>Mamu-A*01</td>
<td>0.020</td>
<td>ann (2009-09-01)</td>
<td>692</td>
<td>87.4</td>
<td>0.749</td>
<td>0.944</td>
<td>0.150</td>
<td>1.597</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>0.021</td>
<td>ann (2007-12-27)</td>
<td></td>
<td>85.8</td>
<td>0.718</td>
<td>0.929</td>
<td>-0.201</td>
<td>1.788</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>0.028</td>
<td>smm (2009-09-01)</td>
<td></td>
<td>85.5</td>
<td>0.712</td>
<td>0.927</td>
<td>-0.001</td>
<td>1.909</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>0.028</td>
<td>smm (2007-12-27)</td>
<td></td>
<td>85.0</td>
<td>0.701</td>
<td>0.920</td>
<td>0.027</td>
<td>1.892</td>
<td>0.014</td>
</tr>
<tr>
<td>Mamu-A*02</td>
<td>0.040</td>
<td>ann (2009-09-01)</td>
<td>249</td>
<td>83.3</td>
<td>0.677</td>
<td>0.930</td>
<td>0.421</td>
<td>2.135</td>
<td>0.197</td>
</tr>
<tr>
<td></td>
<td>0.031</td>
<td>ann (2007-12-27)</td>
<td></td>
<td>82.3</td>
<td>0.646</td>
<td>0.925</td>
<td>0.060</td>
<td>2.322</td>
<td>0.626</td>
</tr>
<tr>
<td></td>
<td>0.084</td>
<td>smm (2009-09-01)</td>
<td></td>
<td>82.3</td>
<td>0.645</td>
<td>0.928</td>
<td>0.100</td>
<td>2.057</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>0.064</td>
<td>smm (2007-12-27)</td>
<td></td>
<td>82.3</td>
<td>0.645</td>
<td>0.918</td>
<td>0.101</td>
<td>2.175</td>
<td>0.046</td>
</tr>
<tr>
<td>Mamu-A*11</td>
<td>0.021</td>
<td>ann (2009-09-01)</td>
<td>307</td>
<td>91.3</td>
<td>0.823</td>
<td>0.974</td>
<td>-0.034</td>
<td>1.532</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>0.018</td>
<td>ann (2007-12-27)</td>
<td></td>
<td>90.5</td>
<td>0.806</td>
<td>0.961</td>
<td>-0.056</td>
<td>1.685</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>0.044</td>
<td>smm (2009-09-01)</td>
<td></td>
<td>89.1</td>
<td>0.778</td>
<td>0.948</td>
<td>0.147</td>
<td>1.908</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>0.054</td>
<td>smm (2007-12-27)</td>
<td></td>
<td>88.6</td>
<td>0.767</td>
<td>0.951</td>
<td>0.129</td>
<td>1.864</td>
<td>0.069</td>
</tr>
<tr>
<td>Mamu-B*17</td>
<td>0.004</td>
<td>ann (2009-09-01)</td>
<td>589</td>
<td>88.6</td>
<td>0.763</td>
<td>0.953</td>
<td>0.104</td>
<td>1.101</td>
<td>0.095</td>
</tr>
<tr>
<td></td>
<td>0.091</td>
<td>ann (2007-12-27)</td>
<td></td>
<td>71.3</td>
<td>0.403</td>
<td>0.798</td>
<td>1.141</td>
<td>2.397</td>
<td>0.476</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>smm (2009-09-01)</td>
<td></td>
<td>83.5</td>
<td>0.858</td>
<td>0.929</td>
<td>0.069</td>
<td>1.392</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>0.003</td>
<td>smm (2007-12-27)</td>
<td></td>
<td>69.8</td>
<td>0.356</td>
<td>0.815</td>
<td>0.350</td>
<td>1.862</td>
<td>0.188</td>
</tr>
</tbody>
</table>

BF: fraction of 9-mer peptides (derived from human and macaque proteomes) that are predicted to bind to HLA-B and Mamu alleles
N: number of available experimental measurements used to test prediction algorithms
ACC: % accuracy of classifying peptides into binders and non-binders (Equation B.1)
MCC: Matthews correlation coefficient (Equation B.2)
AUC: area under the ROC curve
\(\Delta \ln(y)\), \(\sigma\): average bias and error of predicted binding affinity value IC50 (Equation B.6)
Figure B-1: Scatter plots show comparison between experimentally measured and predicted binding affinities of 9-mer peptides to HLA-B*5701 allele (a) and HLA-B*2705 (b). For both alleles the best predictive algorithm (Table B.1) was used. Green data points correspond to measurements, which report exact binding affinity. Red data points correspond to measurements, which report that IC_{50} is larger than that corresponding to its value on the abscissa. Solid lines represent threshold value 500 nM, which divides binder and non-binder peptides. Dashed lines would represent perfect match between predicted and experimentally measured binding affinities. The numbers reported in each quadrant correspond to the number of displayed data points. These numbers are used to calculate accuracy (ACC) and Matthews correlation coefficient (MCC). (c) and (d) show the ROC curves for the best predictive peptide binding algorithms for HLA-B*5701 and HLA-B*2705.
Figure B-2: The predictive algorithms for HLA-B*0801 are not very accurate (see also Table B.1). Scatter plots show comparison between experimentally measured and predicted binding affinities of 9-mer peptides to HLA-B*0801 allele for two predictive algorithms: smm (2009-09-01) on left and smm (2007-12-27) on right. Green data points correspond to measurements, which report exact binding affinity. Red data points correspond to measurements, which report that IC$_{50}$ is larger than that corresponding to its value on the abscissa. Solid lines represent threshold value 500 nM, which divides binder and non-binder peptides. Dashed lines would represent perfect match between predicted and experimentally measured binding affinities. Newer algorithm (a) on average tends to overestimate IC$_{50}$ value, which results in predicting fewer peptide binders. Older algorithm (b) on average tends to underestimate IC$_{50}$ value, which results in predicting more peptide binders. The numbers reported in each quadrant correspond to the number of displayed data points. These numbers are used to calculate accuracy (ACC) and Matthews correlation coefficient (MCC). (c) and (d) show the ROC curves for the two predictive peptide binding algorithms for HLA-B*0801: smm (2009-09-01) on left and smm (2007-12-27) on right.
Figure B-3: The predictive algorithms for HLA-B*3501 are less accurate than that for HLA-B*5701, HLA-B*0702, and HLA-B*2705 (see also Table B.1). (a) Scatter plot show comparison between experimentally measured and predicted binding affinities of 9-mer peptides to HLA-B*3501 allele for the most accurate predictive algorithm ann (2007-12-27). Green data points correspond to measurements, which report exact binding affinity. Red data points correspond to measurements, which report that IC$_{50}$ is larger than that corresponding to its value on the abscissa. Solid lines represent threshold value 500 nM, which divides binder and non-binder peptides. Dashed lines would represent perfect match between predicted and experimentally measured binding affinities. The algorithm on average tends to overestimate IC$_{50}$ value, which results in predicting a smaller peptide binding fraction than reality. The numbers reported in each quadrant correspond to the number of displayed data points. These numbers are used to calculate accuracy (ACC) and Matthews correlation coefficient (MCC). (b) shows the ROC curve for the best predictive peptide binding algorithm for HLA-B*3501.
Appendix C

Host–pathogen interaction dynamics

In this Appendix, we describe in details the dynamical host–pathogen interaction model that was briefly described in Chapter 4 (Figure 4-2). We present also a simplified model of host–pathogen interactions (Section C.2) and show that qualitative results about the effects of cross-reactivity are robust to variations in parameters and model assumptions (Section C.3).

C.1 Host–pathogen interaction dynamics model

We constructed a small model of HIV with distinct epitopes and sequence diversity, based in part on models developed previously (Nowak et al., 1995; Wodarz and Thomsen, 2005). The virus is modelled as displaying $L$ epitopes, each consisting of $M$ amino acid residues that may be of $N$ types. Different viral strains arise through point mutations at the amino-acid sites, giving $(N^M)^L$ distinct strains. The number of different pMHC types is $L \times N^M$, because peptide sequences at epitope positions $1 \ldots L$ are considered to be distinct. The system of ordinary differential equations corresponding to the model in Figure 4-2 and based on previous work (Wodarz and
Thomsen, 2005) is as follows:

\[
\frac{dV_n}{dt} = k_i^n I_n - k_c V_n + k_m \sum_{n:m} (V_m - V_n), \tag{C.1}
\]

\[
\frac{dI^t}{dt} = k_b - k_d I^t - k_t I^t \sum_n V_n, \tag{C.2}
\]

\[
\frac{dI_n}{dt} = k_t V_n I^t - k_d I_n - \sum_i \sum_j \sigma_{i,j} k_P P_{n,j} T_i^*, \tag{C.3}
\]

\[
\frac{dP_{n,j}}{dt} = k_s I_n - k_0 P_{n,j} - \frac{dI_n^{(kill)}}{dt} P_{n,j}, \tag{C.4}
\]

\[
\frac{dP_{n,j}^{APC}}{dt} = k_s' I_n - k_0' P_{n,j}^{APC}, \tag{C.5}
\]

\[
\frac{dT_i}{dt} = -k_a T_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{APC}, \tag{C.6}
\]

\[
\frac{dT_i^0}{dt} = -k_p T_i^0 + k_a T_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{APC} + k_{ra} M_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{APC}, \tag{C.7}
\]

\[
\frac{dT_i^m}{dt} = 2k_p T_i^{(m-1)} - k_p T_i^m, \tag{C.8}
\]

\[
\frac{dT_i^*}{dt} = 2k_p T_i^{(D-1)} - k_d T_i^* - k_m T_i^*, \tag{C.9}
\]

\[
\frac{dM_i}{dt} = k_m T_i^* - k_{dm} M_i - k_{ra} M_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{APC}. \tag{C.10}
\]

Target CD4\(^+\) T cells, \(I^t\), are infected by free virus particles, where \(V_n\) denotes virions of strain \(n\). \(I_n\) denotes CD4\(^+\) T cells infected by virus of strain \(n\), \(P_{n,j}\) is a pMHC complex of peptide \(j\) derived from viral strain \(n\), displayed on the surface of the infected cell, \(P_{n,j}^{APC}\) is a pMHC displayed by APCs and \(T_i\) is a naive CD8\(^+\) T cell of clonotype \(i\). Activated T cells undergo \(D\) rounds of cell division before becoming effector CTLs; \(T_i^0\) is an activated CD8\(^+\) T cell of type \(i\) that has not yet begun dividing and \(T_i^m\) are the dividing cells, where \(m\) runs from 1 to \(D - 1\). Effector CTLs, \(T_i^*\), differentiate into memory CD8\(^+\) T cells, \(M_i\), which are activated upon re-exposure to pMHC.
If T-cell clone \( i \) recognizes pMHC \( j \), \( \sigma_{i,j} \) is 1, and 0 otherwise. In Equation (C.1), \( \sum_{m:m} \) denotes the sum over viral strains \( m \) that are Hamming distance 1 away from strain \( n \). That is, only point mutations are allowed. The third term of Equation (C.4) ensures that if an infected cell is killed, the pMHC bound on its surface must also disappear; \( \frac{df_{k}(\text{kill})}{dt} \) denotes the third term of Equation (C.3), which describes killing of an infected cell by CTLs that recognize pMHC on its surface. Simulations were performed using ode45 and ode15s solvers in MATLAB. A further dynamic model (Section C.2), which does not incorporate target cell limitation and allows unlimited expansion of activated CTLs, was also developed to show robustness of our results to model assumptions (Section C.3).

Rate constants used in the models are given in Table C.1, and are in keeping with values reported in the literature. We assume a concentration of \( 10^6 \) CD4\(^+\) T cells per ml blood before infection, with 1\% of these cells activated and thus initial targets for HIV infection (Sachsenberg et al., 1998; Stafford et al., 2000). The initial conditions of infection in the simulations were one infected CD4\(^+\) T cell per ml of plasma and a naive-CD8\(^+\) repertoire size of one cell per ml of each clonotype. We assume that the number of epitopes, length of each epitope, and number of amino acids \( (L, M, N) \) are all 2, giving 8 pMHC types and 16 possible viral strains. The number of CD8\(^+\) clonotypes was chosen to be 20.

The interplay between antigen and immune receptor diversity is captured in this model through variability in \( \sigma_{i,j} \) and viral fitness. Different fitness levels for different strains of the virus are modelled by randomly selecting \( k_n \), the virus proliferation rate, for each strain from a uniform distribution between 0 and 2,000 per day (Althaus and de Boer, 2008; Parera et al., 2007), with the assumption that the infecting strain has the maximum fitness. The matrix \( \sigma_{i,j} \) encodes the ability of T cells to recognize pMHCs. We generate \( \sigma_{i,j} \) in such a way as to mimic the results of the thymic selection model (Figure 4-1b), to investigate the effects of those predictions on host–pathogen dynamics. That is, we assume that T-cell repertoires restricted by different HLA types differ in the interaction free energies of their TCR–pMHC contacts, and generate \( \sigma_{i,j} \) accordingly using a type of random-energy-like model (Figure C-1). The interaction
Table C.1: Parameters of host-pathogen interaction model shown in Figure 4-2:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
<th>Units</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial target cell</td>
<td>$I'(t = 0)$</td>
<td>$3 \times 10^4$</td>
<td>cell ml$^{-1}$</td>
<td>(Sachsenberg et al., 1998)</td>
</tr>
<tr>
<td>concentration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum virus replication</td>
<td>$k_v$</td>
<td>2000</td>
<td>virions (cell day)$^{-1}$</td>
<td>(Ramratnam et al., 1999; Ribeiro, 2007)</td>
</tr>
<tr>
<td>Virus clearance</td>
<td>$k_c$</td>
<td>20</td>
<td>day$^{-1}$</td>
<td>(Ramratnam et al., 1999)</td>
</tr>
<tr>
<td>Mutation rate</td>
<td>$k_m$</td>
<td>$2.2 \times 10^{-5}$</td>
<td>mutations (base cycle)$^{-1}$</td>
<td>(Huang and Wooley, 2005)</td>
</tr>
<tr>
<td>Target cell production</td>
<td>$k_b$</td>
<td>1000</td>
<td>(cell day)$^{-1}$</td>
<td></td>
</tr>
<tr>
<td>Target cell death</td>
<td>$k_d$</td>
<td>0.1</td>
<td>day$^{-1}$</td>
<td></td>
</tr>
<tr>
<td>Target cell infection</td>
<td>$k_t$</td>
<td>$6.5 \times 10^{-7}$</td>
<td>ml (virus day)$^{-1}$</td>
<td>(Stafford et al., 2000)</td>
</tr>
<tr>
<td>Infected cell death</td>
<td>$k'_d$</td>
<td>0.15</td>
<td>day$^{-1}$</td>
<td>(Bonhoeffer et al., 2003)</td>
</tr>
<tr>
<td>Presentation of pMHC on</td>
<td>$k_s$, $k'_s$</td>
<td>10</td>
<td>day$^{-1}$</td>
<td></td>
</tr>
<tr>
<td>infected cells, APCs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peptide off-rate</td>
<td>$k_0$, $k'_0$</td>
<td>1</td>
<td>day$^{-1}$</td>
<td>(Peter et al., 2001)</td>
</tr>
<tr>
<td>Activated CD8$^+$ expansion</td>
<td>$k_p$</td>
<td>3</td>
<td>day$^{-1}$</td>
<td></td>
</tr>
<tr>
<td>Rate of CTL activation/infected cell killing</td>
<td>$k_a / k_k$</td>
<td>$4 \times 10^{-6}$</td>
<td>ml (cell day)$^{-1}$</td>
<td>(Murali-Krishna et al., 1998)</td>
</tr>
<tr>
<td>Memory cell activation</td>
<td>$k_{ra}$</td>
<td>$8 \times 10^{-6}$</td>
<td>ml (cell day)$^{-1}$</td>
<td></td>
</tr>
<tr>
<td>Effector CD8$^+$ cell death</td>
<td>$k_{dt}$</td>
<td>0.5</td>
<td>day$^{-1}$</td>
<td>(de Boer et al., 2001)</td>
</tr>
<tr>
<td>Differentiation of effector to memory cell</td>
<td>$k_m$</td>
<td>0.008</td>
<td>day$^{-1}$</td>
<td></td>
</tr>
<tr>
<td>Memory cell death</td>
<td>$k_{dm}$</td>
<td>0.015</td>
<td>day$^{-1}$</td>
<td>(Ladell et al., 2008)</td>
</tr>
</tbody>
</table>
Figure C-1: Random energy-like model for generating $\sigma_{i,j}$, the matrix describing recognition of pMHCs by CD8+ T cells. The degree of cross-reactivity in the simulation depends on the uniform distribution from which interaction strengths between individual epitope residues and the TCRs are randomly selected (right). A higher upper limit of the pairwise distribution corresponds to a higher mean and broader distribution of the overall (summed) TCR-pMHC interaction strengths (left). As recognition is considered to occur above a threshold, a broader distribution results in more frequent recognition of pMHCs by T cells in the model, and thus higher cross-reactivity.

The free energy between a T cell and an epitope is given by $\sum_a J(i, j_a)$, where $J(i, j_a)$ is the interaction free energy between T cell of clonotype $i$ and residue $a$ on epitope $j$. Similar to the models used for thymic selection, the total interaction free energy is taken to be the sum of the individual residue interactions and recognition is said to occur when it exceeds a recognition threshold (in the dynamic model, T-cell sequences are not specified explicitly). $J(i, j_a)$ is a random variable chosen from a uniform distribution, and the width of the distribution determines the probability that the summed interaction energy falls above the threshold, and thus the probability that a peptide is recognized by a given T cell. Repertoires generated in this way approach a Gaussian distribution of interaction energies, and the distribution shifts and thus
cross-reactivity increases when the uniform distribution from which \( J(i, j_a) \) is selected is wider. Generating \( \sigma_{i,j} \) in this way allows us to describe variable cross-reactivities of the T-cell repertoire (both intra- and interepitope), and also accounts for correlated interaction energies and thus recognition probabilities of similar peptide sequences.

## C.2 Host–pathogen interaction dynamics for simplified model

The following dynamical model is similar to that in Figure 4-2, but is without the effects of target cell limitation, finite CTL expansion, and CD8\(^+\) memory. Similar models have been studied previously (Handel and Antia, 2008). The following equations describe the model (schematic in Figure C-2):

\[
\frac{dV_n}{dt} = k^n_i I_n - k_v V_n + k_m \sum_{n:m} (V_m - V_n), \quad (C.11)
\]

\[
\frac{dI_n}{dt} = k_t V_n I_n^t - k_d I_n - \sum_i \sum_j \sigma_{i,j} k_k P_{n,j} T_i^*, \quad (C.12)
\]

\[
\frac{dP_{n,j}}{dt} = k_s I_n - k_0 P_{n,j}^t - \frac{dI_n^{(kill)}}{dt} \frac{P_{n,j}}{I_n}, \quad (C.13)
\]

\[
\frac{dP_{n,j}^{APC}}{dt} = k_s I_n - k_0 P_{n,j}^{APC}, \quad (C.14)
\]

\[
\frac{dT_i}{dt} = -k_a T_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{APC}, \quad (C.15)
\]

\[
\frac{dT_{i}^*}{dt} = -k_a T_i \sum_{n,j} \sigma_{i,j} P_{n,j}^{APC} + k_p T_i^*. \quad (C.16)
\]

Rate constants for this model are given in Table C.2, where rate constants governing virus and CD4\(^+\) dynamics are generally adopted from the literature. Approximate rate constants for virus and CD4\(^+\) cell turnover are available from studies in which patient viral loads were perturbed by antiretroviral treatment or plasma apheresis, and the data were fit by dynamical models (Ho et al., 1995; Ramratnam et al., 1999;
Wei et al., 1995). Predicted rate constants for infected CD4+ cell death range from about 0.1 to 1 day^{-1} (Bonhoeffer et al., 2003). This rate constant accounts for cell death due to virus cytotoxicity as well as clearance by effector CTLs and antibodies, and thus is considered an upper bound for $k_d$ in our model, which describes infected cell death by means other than CTL killing. Estimates for the percentage of infected cell death attributable to the CTL response range from 10% to 90% (Althaus and de Boer, 2008; Asquith et al., 2006). Constants for reactions involving CD8+ cells are chosen to give realistic peak and setpoint (in the model described in the main text only) viral loads. The mutation rate from the literature in units of mutations
Table C.2: Parameters of simplified host-pathogen interaction model shown in Figure C-2 (parameters not listed are the same as in Table C.1):

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target cell concentration</td>
<td>$I^1$</td>
<td>$10^4$</td>
<td>cell ml$^{-1}$</td>
</tr>
<tr>
<td>Infected cell death</td>
<td>$k_d$</td>
<td>0.1</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>Presentation of pMHC on infected cells, APCs</td>
<td>$k_s$, $k_s'$</td>
<td>800</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>Peptide off-rate</td>
<td>$k_0$, $k_0'$</td>
<td>40</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>Activated CTL expansion</td>
<td>$k_p$</td>
<td>0.2</td>
<td>day$^{-1}$</td>
</tr>
<tr>
<td>Rate of CTL activation/infected cell killing</td>
<td>$k_a / k_k$</td>
<td>$6 \times 10^{-5}$</td>
<td>ml (cell day)$^{-1}$</td>
</tr>
</tbody>
</table>

(base cycle)$^{-1}$ is converted to $\sim 0.22/(L \times M)$ mutations (amino acid day)$^{-1}$ using an estimate of 1 day for a replication cycle (Perelson et al., 1996) and $\sim 10^4$ base pairs for the size of the virus. In the chronic infection model, the number of cell divisions ($D$) is taken to be 8 (Wodarz and Thomsen, 2005).

If the parameters in the model are chosen such that the virus is able to take hold and expand (Perelson and Nelson, 1999), the qualitative results related to the effects of cross-reactivity are insensitive to the choice of rate constant parameters. This is demonstrated in Figure C-7 for 100-fold variation of the rate constant governing T cell activation, and results were found to be similarly insensitive to variations in the other rate constants (data not shown).

C.3 Robustness of the effects of cross-reactivity to variations in parameters and models.

Figures in this section demonstrate that the qualitative results about the effects of cross-reactivity are robust to variations in parameters and model assumptions.
Figure C-3: Simulation results using the simplified model. HIV viral loads versus time for different cross-reactivities (CR) of the CD8$^+$ T cell repertoire, corresponding to the model in Figure C-2. The black curve corresponds to highly cross-reactive case and the red curve corresponds to lower cross-reactivity. Each curve is averaged over 500 simulations (each simulation represents a person).
Figure C-4: As in Figure 4-4, but for the simplified model (schematic in Figure C-2). When more clones recognize the infecting and emerging strains (left, bottom), the emerging mutant strain (green) is kept in check (left, top). However, when cross-reactivity is low, the likelihood that the mutant strain goes unrecognized is higher (bottom, right), and the mutant strain achieves a large percent contribution of the total virus population (top, right).
Figure C-5: Anticorrelation of simulated peak viral loads with percent contribution of the dominant epitope to the total CTL response for the model in Figure C-2. Percent contribution is calculated as the number of activated CTLs recognizing the immunodominant epitope over the total number of activated CTLs in the simulation. The immunodominant epitope is defined as the epitope recognized by the largest number of CTL clones. Lower percent contributions are achieved when the CD8+ T cell repertoire is less cross-reactive, which also correlates with higher viral loads, as found experimentally by Altfeld and coworkers (Streeck et al., 2009). The black points and bars correspond to the average and standard deviation of 500 simulations for each level of T cell cross-reactivity, with the level of cross-reactivity increasing from left (probability of 0.28 that a given epitope is recognized by a particular CTL) to right (probability 1). Varying other parameters in the model, including peptide presentation rate, does not capture this behavior (Figure C-11).
Figure C-6: Peak viral load versus average number of CD8\(^+\) clones recognizing a pMHC in each simulation in the simplified model (schematic in Figure C-2). “Escape” is taken to mean that the population size of a mutant viral strain has become larger than that of the infecting strain at some point during the simulation time (0 to 80 days). As exemplified in Figure 4-4, the smaller the number of clones recognizing each pMHC (corresponding to lower cross-reactivity), the higher the chance of escape.
Figure C-7: Insensitivity of qualitative results to changes in CD8⁺ T cell activation rate for the simplified model (schematic in Figure C-2). Left panels show simulation results with varying cross reactivity for activation rate 10kₐ (kₐ given in Table C.2), while right panels show results for kₐ/10. Insensitivity of qualitative results to parameter variation was found for other rate constants in the model also (not shown).
Figure C-8: Anticorrelation of simulated viral loads with percent contribution of the dominant epitope to the total CTL response, as in Figure C-5, but corresponding to the model described in Chapter 4 and Section C.1 (Figure 4-2). Viral load and % contribution were calculated at day 200 in the simulations, to approximate viral load setpoint. Both models give qualitatively similar results. Thus, the result that a cross-reactive repertoire results in low viral loads and high % contribution of responses to the dominant epitope is insensitive to the choice of dynamical model.
Figure C-9: As in Figure C-6, but for the model described in Chapter 4 and Section C.1 (Figure 4-2). As the number of clones recognizing a pMHC increases, the setpoint viral load and probability of escape decrease.
Figure C-10: Insensitivity of qualitative results to changes in CD8+ T cell activation rate for the model described in Chapter 4 and Section C.1 (Figure 4-2). Left panels show simulation results with varying cross reactivity for activation rate $10k_a$ ($k_a$ given in Table C.1), while right panels show results for $k_a/10$. The setpoint viremia level depends strongly on $k_a$, but the qualitative correlation between viral load and % contribution of the immunodominant epitope (a and b) and the number of clones targeting a given pMHC (c and d) is the same. The same insensitivity of qualitative results to parameter variation was found for other rate constants in the model also (not shown). Note that for a lower activation rate (right panels), the probability of escape is reduced (fewer green points), because of the lower overall immune pressure exerted by the same number of T cells.
Figure C-11: Setpoint viral load versus % contribution of immunodominant epitope, as in Figures C-5 and C-8, but where the rate of peptide presentation (not cross-reactivity), is varied. Points correspond to $k_s$ values of 200, 100, 40, 20, 10, 5, and 1 (day$^{-1}$), with faster presentation of pMHC corresponding to reduced peak viral loads. One potential effect of B57 binding fewer peptides is that the cell-surface concentration of immunogenic peptides could increase, because competition with other peptides for binding to MHC would be reduced. This would then be an additional mechanism for control of viral load. Increasing $k_s$ has the effect of raising cell-surface concentration of pMHC and reducing viral load. As the figure shows, varying only this parameter leads to correlation of high % contribution with high viral loads, in contrast to the result of Altfeld and coworkers (Streeck et al., 2009). Variation of other rate constants in the model gave similar results or had no effect on % contribution (not shown). Therefore, only varying the cross-reactivity recapitulates the experimental results. The issue of peptide presentation could be important if HLA molecules like HLA-B*5701 presented far fewer HIV epitopes and so, due to less competition, these epitopes were presented faster, and hence, in greater amounts. We have used the predictive algorithms and the published HIV proteome (HXB2) to estimate the number of HIV epitopes that can bind to the alleles we have identified from our data (Figure 4-5) to be associated with control or progression. Approximately 40 peptides can bind to HLA-B*5701 and HLA-B*0702 and approximately 60 peptides can bind to HLA-B*2705 and HLA-B*3501. Thus, the number of HIV peptides that can bind to these alleles does not correlate with disease outcome.
Bibliography


