8

A Monte Carlo Simulation for the Construction of Cytotoxic T Lymphocytes Repertoire

Filippo Castiglione
Istituto per le Applicazioni del Calcolo "M. Picone" (IAC)
Consiglio Nazionale delle Ricerche (CNR)
Italy

1. Introduction

The immune system of vertebrate living beings is a very complicated system that has evolved a set of mechanisms to get rid of potential pathogens he get in contact with. These mechanisms are so finely tuned that we don't even realize how much work our immune defenses are carrying out each moment to keep us disease free. Unfortunately, as one can figure, the whole system is not error free.

For example, autoimmune diseases arise when at least one of the mechanisms meant to preserve *tolerance of self* breaks. The immune self tolerance is the process by which the immune system refrain from attaching the host own body. This is normally the norm but given the huge complexity of the interdependencies among immune components (cells, molecules, organs, signals, etc.), and also given that the immune defenses are not static but rather dynamic and ever changing during our lifetime, it should not be surprising to learn of the different autoimmune diseases known to date.

Tolerance is the evolutionary result of a multi-layer system whose goal is to weed out self-reactive cells. Of these mechanisms, lymphocytes T education in the thymus organ represents the very first one.

In this article we describe a Monte Carlo method to simulate the maturation of key immune cells. Before describing the algorithm we give a brief introduction of how the immune system works.

The goal of the present work is purely methodological. In fact, while we focus on the specific aspect of the examination phase of the T lymphocytes maturation, we use immuno-informatics data and methods to perform computer simulation of the whole process without taking into account, for example, the anatomical structure of the thymus organ where the process of lymphocytes education takes place. Moreover, for reasons that will be mentioned later, we will specifically deal with $CD8^+$ T-cells (i.e., CTLs) selection rather than on $CD4^+$ T-cells (T helpers).

1.1 About the immune system

The immune system is the sum of a number of functions exerted at different spatial scales, from the micro of molecules to the macro of tissues and organs. White blood cells like the lymphocytes and the phagocytic/dendritic cells are the most important. Their action is to seek out and destroy disease-causing organisms or substances (generically called antigens)

like bacteria or virus and to mount a response in terms of production of antibodies or antigen-specific cells. These are the two types of immune response, respectively humoral and cytotoxic. Moreover, talking about immunity we have to distinguish between the innate immunity and the acquired immunity. The first consists of the basic protection mechanisms like the skin/mucosa barriers and the phagocytic *aspecific* activity of certain kinds of cells. The second type of immunity, the adaptive, is the most recent in evolutionary terms, and it is what confers the immune system its adaptability to new invaders which are in turn the result of natural selection pressures.

While the cellular components of the innate immunity recognize structures shared by different classes of microbes, which is as to say that its recognition ability is *hard wired* into the membrane receptors, the adaptive immunity consists of a multitude of *clones* bearing different cell receptors and therefore able to recognize different antigens. The adaptability of this system comes from the fact that the immune response elicited by a pathogen is *specific*: only those clones able to recognize the antigen will start to proliferate, creating an army of cells with tailored weapons against that specific antigen.

A crucial step in the recognition of the antigen by the lymphocytes is its *presentation* by certain types of cells called *antigen presenting cells*. These cells capture the antigen, digest it, and then show, on their membrane surface, bits and pieces of the antigen attached to a molecules called *major histocompatibility complex* (MHC). The lymphocytes through their receptors, can only see the antigen when attached to the MHC.

An important feature of the immune system is its ability to remember already encountered pathogens. Memory resides in antigen-specific cells that live much longer than normal, conferring the system the ability to mount a more effective and swift responses the second time it is threaten by the same pathogen. Memory is the reasons why vaccines can confer long-lasting protection against infections.

1.2 The diversity of lymphocytes receptors

Without going into detail on the mechanisms of generation of receptor diversity (combinatorial and junctional diversification), we note that somatic recombination may result in potentially more than 10^{15} receptors with different specificities. Given this potential diversity, it is estimated that only 10^7 are actually expressed receptors; i.e., different clones circulating in our body every day.

This enormous diversity of receptors occurs during the maturation process of lymphocytes. This process starts from bone marrow stem cells and includes three phases: the proliferation of immature cells, expression of the antigen receptor genes and selection of lymphocytes that express an antigen receptor *useful*. Lymphocytes that are present in the immune system are those who have passed the ripening process. The maturation of T lymphocytes occurs in the thymus and during this phase are the receptors that will recognize any foreign antigens to the body. Since the process of gene recombination is random, thymus deletes those cells whose receptors have high specificity to proteins of the host, while those who are selected will be able to recognize the MHC molecules of the host. This is crucial because the TCR must also recognize residues on the MHC molecule in order to make the recognition of an antigen. A similar process occurs for B cells in bone marrow but mature and not need to recognize MHC molecules.

1.3 Immune effector mechanisms

There are two types of cell-mediated reactions deputies to remove different types of intracellular microbes: the first $CD4^+$ T cells activate macrophages so that they destroy

the microbes contained in their vesicles, the other $CD8^+$ lymphocytes that kill cells micro-organisms in their cytoplasm, eliminating the reservoir of infection. The humoral immunity is mediated by antibodies and instead represents the arm of acquired immunity deputy to neutralize extracellular microbes. Antibodies are molecules of the family of immunoglobulin. Organism are produced five different classes of Ig that are differentiated by their heavy chain constant region determines the effector function.
A B lymphocytes activated by antigen recognition may differentiate into a plasma cell or a cell that produces antibodies. These have the same specific receptor that recognized the antigen but are able to act remotely as soluble. The antibodies act by using the antigen binding region to bind to microbes and toxins by blocking the pathological effects, while using the constant portion of the heavy chain to activate different effector mechanisms that cause the deletion. Effective action can only be made when several antibodies recognize an antigen and bind to it.

1.4 The components of the immune system

Lymphocytes are cells with receptors specific for the antigen and are the central component of acquired immunity. Very similar, the lymphocytes are very heterogeneous in terms of functional and phenotypic.
B lymphocytes: are the only cells that produce antibodies and mediate humoral immunity. Express on their surface antibody molecules that serve as receptors for antigen recognition and to start the activation process. Soluble antigens or bound onto the surface of bacteria or other cells can bind to these receptors, initiating the humoral response.
T Lymphocytes: mediate cellular immunity. Their receptor recognizes antigen peptides only fragments linked to protein molecules specialized in presenting antigen (MHC I or MHC II). The most important T-cells are: or T helper (CD4 + or): Their TCR recognizes peptides bound to MHC class II. Their function is to help B cells produce antibodies and phagocytes to destroy microbes incorporated by cytokine release.
Cytotoxic T lymphocytes (CTL or CD8 + or): Their TCR recognizes peptides bound to MHC class I. Their function is to kill cells infected by intracellular microbes.
Dendritic cells: Although in principle are cells of innate immunity, play their most important function in presenting antigens to compartment specific immunity. Dendritic cells capture antigens that penetrate through the epithelium and transport them to draining lymph nodes. Lymph nodes expose their membrane fragments of microbial protein antigens to activate T lymphocytes with the specific receptor.
The various cells cooperate to protect the body from infections and illness. They can communicate through chemical mediators to orchestrate the response; they circulate throughout the body in the lymphatic and blood system to patrol every single organ. Lymphocytes are equipped with a transmembrane molecule called receptor that are used to bind to antigens. This binding event is the first step in recognizing anything dangerous. It is therefore clear that mother nature has constructed these membrane receptor very carefully. In fact, if the lymphocytes recognize a *self* molecule, soon the immune system would start to destroy it. A self molecule is so called because it belongs to our own body cells, therefore being the target of an immune attack causes inflammation and damage and leads to autoimmune disorders.
The causes of autoimmune diseases are unknown, although it appears that in many cases there is an inherited predisposition to develop them. In a few types of autoimmune disease (such as rheumatic fever), a bacteria or virus triggers an immune response, and the antibodies

or T-cells attack normal cells because they have some part of their structure that resembles a part of the structure of the infecting microorganism.
The membrane receptors of lymphocytes are randomly arranged so, potentially, they can bind to any molecule. How is then tolerance achieved?
One mechanism is to scrutinize the immature lymphocytes for potentially autoreactive ones and eliminating them before they go into circulation. For lymphocytes T helpers and cytotoxic T cells, this step is performed in the organ from which these cells take their name; the thymus. Thymus education represents the very first method to limit autoreactivity.

1.5 The thymus organ

The functioning of the thymus is far from being fully understood. What is surely known is that bone marrow-derived T lymphocytes that do not yet express co-receptor (called double negative DN or $CD4^{-}CD8^{-}$), enter the thymus, migrate to the thymic cortex (the outer region) and proliferate. Most of them also begin to express both CD4 and CD8 co-receptors (double positive DP, or $CD4^{+}CD8^{+}$) together with the T-cell receptor molecule (TCR) and its associated accessory proteins (the CDR3 protein complex). At this stage the immature lymphocytes express high levels of Fas antigen which can trigger death when ligated and produces very little Bcl-2, a cellular protein that protects against apoptosis. This means that they are very sensitive to signals that can trigger death by apoptosis. These apoptotic signals will come from antigen presenting cells (APCs). In particular, in the thymic cortex they interact with cortical thymic epithelial cells and maturate in single positive (SP) cells ($CD4^{+}CD8^{-}$ or $CD4^{-}CD8^{+}$). Then they migrate to the medulla and undergo a series of interaction with thymic dendritic cells and medullary thymic dendritic cells. This "walk" inside the thymus lasts for about 2 weeks. Those lucky thymocytes who survive the selection leave the organ to patrol the body in the search for potential pathogenic agents (Murphy et al., 2008).
The interactions with APCs, who act as "examiners", is meant to score the cells according to the ability to recognize the major histocompatbility complex (MHC) molecules (this is called *MHC restriction*), and also to the inability to bind peptides that belong to "self" (this is the *tolerance induction*). These two requirements are important to guarantee that a matured T lymphocyte is fully functional and, at the same time that, should he be able to recognize self molecules, he will not leave the thymus to cause damage.
APCs score T-cells through their T-cell receptor, a membrane protein whose extra cellular domain is able to bind the MHC-peptide complexes. In contrast to B cells, T-cells cannot recognize a pathogen on its own since they need to be presented in the context of an MHC-peptide complex. For most T-cells, the TCR is a heterodimer, composed of an α and a β chain. There is another $\delta\gamma$ based TCR which is seldom encountered. The TCR belongs to the immunoglobulin superfamily and have one N-terminal immunoglobulin variable (V) domain, one constant (C) domain, a transmembrane domain and a short cytoplasmic C terminal segment. The variable region on the TCR is potentially unique for each T-cell, and is composed of three parts on both the α and β chains, called complementary determining regions (CDRs). CDR3 is thought to be the main molecule interacting with the antigen, while CDR2 would be interacting with the MHC molecule. It is important to note that TCRs originate from a limited number of genes (65 V genes, 27 D genes, 6 J genes), but despite this, the immune system is able to engender a great number of receptors (Goldsby et al., 2000; Murphy et al., 2008).
The generation of the TCR is similar to the one of immunoglobulins (BCR) in B cells. The α chain is generated by VJ recombination while the β chain relies on V(D)J recombination. The gene segments are then randomly joined together to produce the final TCR. The CDR3 region

corresponds to the junction of the V and J segment on the α chain and the V D and J β chain, explaining its high variability and its role in antigen binding thereof.
The TCR selection in a primary organ like the thymus is called *central tolerance induction* and is the first mechanism to assure that most auto-reactive cells are eliminated. Fortunately is not the only one. In fact, since this mechanism is not hundred percent efficient, the immune system is equipped with other mechanisms that constitute the *peripheral tolerance* and induce cell death in auto-reactive lymphocytes. If this does not happen, then autoimmune diseases arise.
Understanding the complex machinery of how thymic selection imparts MHC-self-peptide complex restriction and at the same time a high degree of self tolerance on the T-cell repertoire is a very challenging task and a lot of aspects remains unclear (Klein et al., 2009).

1.6 A short review of mathematical models of T-cell development

Few mathematical models have been used to study specific issues of T-cell development. Most of these models are based on ordinary differential equations. For example, one of the first mathematical model to study thymocyte subset dynamics was introduced in (Mehr et al., 1995). In this model the equations define time evolution of thymocyte subsets, including DN, DP, CD4SP, and CD8SP cells. The model predict that negative selection likely operates at the DP stage or later. Moreover the model revealed that the CD4SP over CD8SP cell ratio fits the "instructive" theory of thymic lineage commitment (Germain, 2002; von Boehmer & Kisielow, 1993). In (Mehr et al., 1997) the idea that thymocytopoiesis may be subject to feedback regulation by mature lymphocytes is proposed and experimental data was analyzed using mathematical models. Another equation-based model was used to to compare the intrathymic development of bone marrow precursors, derived either from young or old donors (Mehr et al., 1993). In Mehr et al. (1998) the phenomenon of *MHC-linked syngeneic developmental preference* was analyzed by a mathematical model. In another study, the authors focused on the naïve T-cell compartment defined by the presence of T-cell receptors excision circles formed during T-cell receptor gene rearrangement (Hazenberg et al., 2000).
In contrast to the above mentioned studies, the model introduced in (Efroni et al., 2005) takes into account the spatial information in thymocyte development, by using agent-based modeling. This modeling paradigm is ideal for uniquely identify cellular characteristics, like for example, receptor expression, to distinguish the different stages of the cell cycle in specific anatomical compartments (Efroni et al., 2007; 2003). Another example of a discrete spatial model can be found in (Souza-e Silva et al., 2009). In this study a cellular automaton was constructed to describe thymocyte migration and development in the thymic microenvironment.
The model presented in the present article takes yet another approach. We simulate MHC restriction and tolerance induction by a stochastic model that includes bioinformatics methods to assess the affinity between a cell receptor and an MHC molecule bound to a self peptide. This study follows the lines of Morpurgo *et al.* ((Morpurgo et al., 1995)) but diverges from it in that the molecules represented (i.e., TCRs, MHCs and self peptides) are not strings of zeros and ones (i.e., binary strings) but rather strings of letters representing the twenty amino acids. Moreover, most importantly, the function used to compute the affinity among these molecules is provided by data-driven machine learning bioinformatics methods.
In immunology what is of outmost importance is to "predict" whose part of the antigenic molecule will constitute an immunogenic epitope. Broadly speaking there are two ways of doing it. The first is to simulate the chemical-physical interactions between peptides and

MHC molecules (e.g., NAMD, NAnoscale Molecular Dynamics or ABF, Adaptive Biasing Force software (Darve & Pohorille, 2001)), that takes hours to simulate a single peptide-MHC interaction what in reality lasts fractions of a second. The second possibility is to resort to bioinfomatics approaches that use machine learning and statistical methods to extract and generalize information from available experimental data of MHC-peptide sequences (for a review see e.g., (Lundegaard et al., 2007)). These methods take a fraction of a second to run on common workstations hence, from this point of view, are preferable to the first one.
Immunoinformatics is a new discipline emerging from the growing knowledge gathered for decades in experimental immunology and immunogenomics (Korber et al., 2006; Petrovsky & Brusic, 2002). Being both an experimental and theoretical field, it is foreseen that immunomics will play an important role for the future of immunology (Petrovsky & Brusic, 2006).
The goal of the present work is to use machine learning techniques for molecular-level predictions of major histocompatibility complex-peptide binding interactions (Lund et al., 2004; Nielsen et al., 2007; 2004), and a more general protein-protein potential estimation (Miyazawa & Jernigan, 1999) to perform Monte Carlo simulation of the selection of thymocytes in the thymus.
MHC class I binding predictions methods based on machine learning have increased their accuracy over the years, thus leading to reliable predictions. The same level of predicting power has not yet been reached by class II prediction methods (Lin, Ray, Tongchusak, Reinherz & Brusic, 2008; Lin, Zhang, Tongchusak, Reinherz & Brusic, 2008). This is thought to be partly due to the structure of the MHC molecules, which binding pockets are open in class II, thus allowing peptides of different length to bind to the groove. In contrast, class I molecules restrict the size of the peptides they bind to to 8-12 amino-acids (Yewdell et al., 2003; 1999), with an average length of 9 amino-acids. Therefore for convenience we restrict the attention to the education of $CD8^+$ T-cells rather than on CD4 T-cells (Lund et al., 2005).

2. The construction of the T-lymphocytes repertoire: a Monte Carlo method

The thymus organ is modeled as a simple *two-stage* filter. In the first stage, APCs give a survival signal to immature T-cell if a binding to the MHC-peptide complex occurs but "weakly", i.e., no binding will drive the cell to apoptosis; in the second stage the survival signal is given if the affinity to MHC-peptides presented by APCs is not "too high", i.e., high avidity for self peptides drives the cell to apoptosis (see left panel of figure 1).
In real life, bone marrow-derived T-cells entering the thymus initially home in the *thymus cortex* where they start to proliferate. Shortly after, rearrangement of the gene segments that encode the α and β chain of the T-cell receptor begins. Somatic rearrangement makes the TCRs highly diverse (about 10^8 (Arstila et al., 1999)).
All we need to account in the simulation is this huge variety. We do it by assigning random amino acid string receptors to each lymphocyte. Since the complementarity-determining region (CDR) of TCRs is what interacts directly with antigenic peptides bound to grooves of MHC molecules, the amino acid string we define for each T-cell is meant to represent not the whole TCR but rather its CDR.
The choice for the length of this string, thus the size of the repertoire, is not an easy one. Studies of various T-cell subsets from humans in physiological and pathological conditions have found an average length between 6 and 60 bases with a high variability in the different groups but only for the CDR3 suggesting that the whole CDR is much longer (Nishio et al., 2004). However not all arrangements of the genes gives a functional TCR and therefore for the

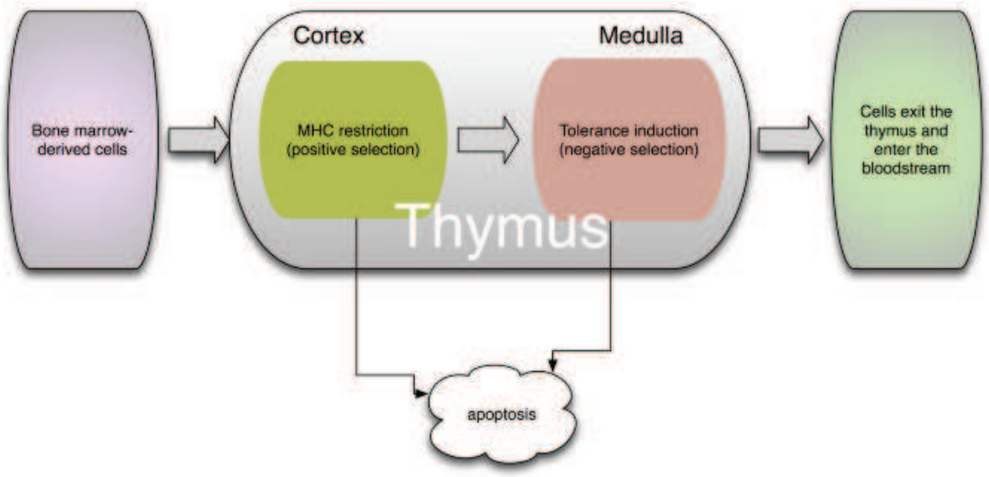

Fig. 1. The two-layer filter realized by the thymus to eliminate auto-reactive T lymphocytes. T-cells develop *self tolerance* during negative selection, whereas they are "discarded" as useless during positive selection.

sake of simplicity we decide to use only 8+12=20 letter to represent respectively the CDR of the α and of the β region of the T-cell receptor.
Self antigens are molecules, products of the organism own cells and as such tolerated by the parent immune system. We represent self molecules as amino acid strings. The question is how long. As we mentioned earlier, the prediction of class I MHC peptides is more accurate than those of class II (Lund et al., 2005). The reason for this is that in contrast to class II MHC that has the ends of the groove open, those of class I MHC are closed so a protein fragment to fit in properly must be about nine amino acid in length. In other words, class I-type epitopes are linear sequences of 8 to 11 amino acids that are processed from any protein of the pathogen. However each MHC class I molecule (whose total number surpasses the thousands of alleles to date (Nielsen et al., 2007)), is characterized by a specific binding motif that is possible to "decode". For the vast majority, the motif length is nine amino acids long, therefore by restricting our study to CTL's tolerance induction, we can define "self" any string of nine amino acids.
To be presented by the APCs, the peptide has to attach to the MHC molecule and therefore we need a method to predict peptide binding to class I MHC.

2.1 Identifying peptides for class I MHC

Methods for class I T-cell epitope prediction rely mainly on machine learning techniques (Lund et al., 2004; Nielsen et al., 2004). In particular neural networks seems to perform well. For example the authors in (Nielsen et al., 2003) show that neural networks trained to predict binding versus non-binding peptides, are superior to other methods (Lin, Ray, Tongchusak, Reinherz & Brusic, 2008; Lin, Zhang, Tongchusak, Reinherz & Brusic, 2008). Furthermore, quantitative neural networks allow the straightforward application of a *query by committee* (QBC) principle, in which particularly information-rich peptides can be identified and subsequently tested experimentally. Moreover, iterative training based on QBC-selected peptides considerably increases the sensitivity of the prediction without compromising its efficiency (Buus et al., 2003).

The Monte Carlo method described herein does not make a direct use of neural network but rather employs a derived *Position Specific Scoring Matrix* (PSSM)-based method. In practice, for each MHC allele we use the *binding motif matrices* generated from the neural network methods as described in (Nielsen et al., 2007). In short, first the neural network is used to rank a set of 10^6 randomly selected natural peptides from the human genome, then the top one percent of the peptides are flagged as binders and used to generate a binding motif, that is a 9 by 20 matrix. These matrices (one for each MHC) are calculated using sequence weights, and are corrected for low counts (Altschul et al., 1997; Nielsen et al., 2004).
Binding motifs matrices are made of *propensities* calculated as $2\log_2(p/q)$, where p is the probability of finding a given amino acid at a given position, and q is the probability of finding that amino acid in any protein in general. These propensities are computed for each of the nine positions on a potential epitope, and give the propensity for each of the 20 amino acids.
Furthermore, we set an allele-specific threshold $\Theta_{\mathbf{H}}$ as the average score of the low-scoring binders in the top one percent of the binders (see (Nielsen et al., 2003; Yewdell et al., 1999) for details).
Finally, having the PSSM and its corresponding threshold $\Theta_{\mathbf{H}}$, we can discriminate binders and non binders by calculating the score and comparing it with $\Theta_{\mathbf{H}}$. More formally, let Ω be the set of amino acid symbols Ω and $\mathbf{p} = [a_1, a_2, \ldots, a_{l(\mathbf{a})}]$, represent a contiguous stretch of amino acids, with $l(\mathbf{p})$ the length of the sequence and $a_i \in \Omega$ the i^{th} amino acid in the sequence. For a given 9-mer $\mathbf{p} = [a_1, a_2, \ldots, a_9]$, the sum of the values at each position in the scoring matrix $\mathbf{H} = \{\mathbf{h}_{a_i,i}\}_{i=1,\ldots,9, a_i \in \Omega}$, of a particular MHC gives the propensity to bind that MHC, i.e.,

$$\mathbf{p} \text{ is a peptide} \iff \sum_{i=1}^{9} \mathbf{h}_{a_i,i} \geq \Theta_{\mathbf{H}}. \tag{1}$$

Therefore of all possible 9-mers only those for which $\sum_{i=1}^{9} \mathbf{h}_{a_i,i} \geq \Theta_{\mathbf{H}}$, where $\Theta_{\mathbf{H}}$ is the allele-specific threshold, are considered epitopes that can be presented by antigen processing cells.
The second thing we need is to have a way to assess whether a TCR interacts with a given MHC-peptide complex or not.

2.2 Interaction with antigen presenting cells

To date, there is no general method that can be used to predict if, for example, a TCR will interact with any given MHC-peptide complex. For this reason we resorted to the Miyazawa-Jernigan residue-residue potential (Miyazawa & Jernigan, 1996) to score the strength of the interaction. The work performed by Miyazawa and Jernigan on protein energy potentials (Miyazawa & Jernigan, 2000) provides us with a method for assessing the chances of direct interactions among proteins in the simulation. The protein-protein potential concept was derived from the analysis of 3D structures in which the relative position of amino acids were determined. The contact potential matrix estimated by Miyazawa and Jernigan reflects the entropy between two residues; a low entropy means that the pair of residues has low energy and therefore that interaction is possible.
For $CD8^+$ T-cell recognition, the procedure to compute the binding score requires the definition of class I MHC specific *contact matrices*. These matrices can be computed by looking at known protein 3D structures found in the Protein Data Bank (PDB, `www.pdb.org`). In the end, we decided to adopt just one MHC contact matrix for *all* class I alleles. The reason for this is twofold: i) we did not find too much differences among contact matrices of different alleles and ii) the one we decide to use has been calculated from the best resolution data (i.e., 1.4 Å).

This contact matrix (that we call **C**) was calculated taking residues that i) are within a distance of 5 Å and, ii) show contacts between the MHC-epitope complex and the two chains (heavy and light) of a bound TCR.
The distance of 5 Å was selected because most crystal structures with experimentally verified B cell epitopes show that the residues on the antibody in contact with an epitope lie within a 5 Å radius. Again the reason for looking at B cell epitopes is twofold: firstly TCRs and immunoglobulins are not that different, and secondly, there are a lot more antigen Ig structures than MHC-pep-TCR structures.
We extend the use of this value to the minimum distance needed between residues for molecular interaction. By using the solved structures, it is possible to determine which residue on a TCR binds to the MHC or to the peptide. The contact matrix derived for class I binding that we use in the simulation is shown in the right panel of figure 1.
Similarly, one determines the residues that are generally in contact with the TCR. These are what we call MHC *pseudo-sequences* (indicated with `MHC`). Again, the contact residues are defined as being within 5.0 Å of the peptide in any of a representative set of HLA-A and -B structures with 9-mer peptides and TCRs (Nielsen et al., 2007).
The contact potential defined between a TCR and an MHC-peptide complex is thus based on the Miyazawa-Jernigan score as follows. Let $\{\mathbf{M}_{a,b}\}_{a,b\in\Omega}$, be the matrix in (Miyazawa & Jernigan, 2000), $\mathbf{x} = [x_1, \ldots, x_{l(\mathbf{x})}]$ a TCR, $\mathbf{y} = [y_1, \ldots, y_{l(\mathbf{y})}]$ a MHC-peptide complex composed by the pseudo sequence of the MHC and the peptide molecules, and **C** the contact matrix. We first compute the binding affinity between **x** and **y** as

$$\hat{M}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{l(\mathbf{x})} \sum_{k=1}^{l(\mathbf{y})} (\mathbf{M}_{x_j, y_k} \cdot \mathbf{C}_{j,k}).$$

Then, since we need to define a probability, this value is normalized and further compared to a threshold value

$$M(\mathbf{x}, \mathbf{y}) = \begin{cases} \frac{\hat{M}(\mathbf{x},\mathbf{y}) - \mu_{\hat{M}}}{k \cdot \sigma_{\hat{M}}} & \text{if } (\hat{M}(\mathbf{x}, \mathbf{y}) - \mu_{\hat{M}})/(k \cdot \sigma_{\hat{M}}) \geq P_{95}, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\mu_{\hat{M}}$ and $\sigma_{\hat{M}}$ are respectively the average and the standard deviation that have been previously estimated, k is a free parameter chosen to have $M(\mathbf{x}, \mathbf{y}) \leq 1$, and P_{95} is, the 95^{th} percentile rank of the estimated distribution. Finally, we use the binding affinity $M(\mathbf{x}, \mathbf{y})$ of eq(2) as the probability of the binding between **x** and **y**.

2.3 Positive selection of TCRs

Given $\mathtt{n_{mhc}}$ major histocompatibility complex molecules, we calculate the probability to pass the positive selection of a cell bearing a random TCR as follows

$$\mathrm{Pr}^{+} = 1 - \prod_{j=1}^{\mathtt{n_{mhc}}} (1 - M(\mathtt{TCR}, \mathtt{MHC_j}\star)), \tag{3}$$

where $M(\cdot, \cdot)$ is the Miyazawa-Jernigan contact potential in eq(2), and $\mathtt{MHC}\star$ indicates that only residues of the TCR in contact with the MHC are taken into account (the wildcard $\star$ means that no matter what the peptide is, we give more weight to the MHC rather than to the peptide by summing a low constant value for each residue).

2.4 Negative selection of TCRs

If the cell survives the previous step with probability given by equation 3, then it can be negatively selected with probability

$$\Pr^{-} = \Big[\prod_{j=1}^{\mathtt{n_{mhc}}}\prod_{k=1}^{\mathtt{n_{self}}}(1 - M(\mathtt{TCR}, \mathtt{MHC_j self_k}))\Big]^{\mathtt{E}} \tag{4}$$

where the $\mathtt{MHC_j self_k}$ is a string composed by the pseudo sequence $\mathtt{MHC}_j$ and the chosen peptide $\mathtt{self}_k$. The parameter $\mathtt{E}$ represents the efficiency in the selection process: higher efficiency means better filtering, that is, less self-reactive cells will slip out the thymus. From the point of view of the calculation of the survival rate of the immature cells entering the thymus, the negative selection is treated as if the thymus were composed by $\mathtt{E}$ sub-layers simulating as many encounters with each thymic cell receptor specificity because of the crowded nature of the thymus.

Equation 4 gives the probability that the TCR does not matches any of the self molecules with any of the MHCs.

Finally, we allow the T-cell to leave the thymus and to reach a secondary organ as a mature thymocyte with a probability given by the product of the probability of being positively selected and the probability of being negatively selected,

$$\Pr(\text{TCR is selected}) = \Pr^{+} \cdot \Pr^{-}.$$

The whole algorithm is summarized as pseudo-code in figure 2.

3. The outcome

Figure 3 shows the Logo plots of the TRC sets that have been filtered in during the thymus selection, that is, those bear by cells leaving the thymus as operational CTLs. The height of the letters reflects the Shannon information at individual positions.

In figure 3 it can be observed that there are specific preferential positions for some amino acids (i.e., smaller entropy of amino acid distributions in Logo plots). The analysis of these positions reveals a smaller entropy (i.e., higher bars) which is consistent with the contact matrix shown in the right panel of figure 1; the more contact positions on the TCR, the less the degrees of freedom.

Interestingly, positions 13 and 19 have respectively no interaction (position 13) or just one interaction (position 19) with the peptide (see right panel of figure 1). This means that negative selection does not influence the amino acid distributions for these positions in figure 3, resulting in similar conservation rates. In contrast, position 16 strongly interacts with both MHC (4 contacts) and peptide (3 contacts) so that it provides a strong constraint on the corresponding TCR residues, resulting in a smaller bar of the Logo plot because of a lack of match. Summarizing, the positive selection step sets a strong constraint to TCR sequences whereas the effect of negative selection provides one more constraint to the sequences at position 16.

Once we have selected the TCRs, $\mathtt{tcr}_1, \ldots, \mathtt{tcr}_N$ by executing the algorithm described in figure 2, we can compute the average auto-reactivity as the average of the probability to recognize at least one self peptide attached to one MHC molecule: $\alpha = \frac{1}{N}\sum_i^N \left(1 - \mathtt{Pr}^{-}\right)$. This value depends on the parameters chosen; for example it depends on $\mathtt{E}$ and on the $\mathtt{MHC}_1, \ldots, \mathtt{MHC}_{n_{mhc}}$. It also depends on the self molecules $\mathtt{self}_1, \ldots, \mathtt{self}_{\mathtt{n_{self}}}$. However, for $\mathtt{n_{self}}$ large enough, the actual amino acid strings are less important their number $\mathtt{n_{self}}$ itself. This

```
input(E, N, n_mhc);                                          input parameters
for (i = 1; i≤ n_mhc; i++) {
        read(H_i);                                           read the MHC specific matrix
        read(MHC_i);                                         read the MHC pseudo sequence
        read(Θ_{H_i});                                       read the MHC specific threshold
}
input(n_self);                                               input the number of self molecules
i == 1;
while i ≤ n_self {                                           randomly choose n_self 9-mers
     p_i = random();                                         pick randomly a self peptide
     if ( ∃k : binds(H_k, p_i) == true) {                    if it binds at least one MHC molecule than accept it
          Self ← p_i;                                        p_i will be shown as self
          i++;
     }
}
i=1;
while i ≤ N {                                                select TCRs
     TCR_i = random();                                       generate a random amino acid string for TCR
     Pr^+ = compute(TCR_i);                                  compute Pr^+ as in equation 3
     Pr^− = compute(TCR_i);                                  compute Pr^− as in equation 4
     if (rand(0,1)<Pr^+ · Pr^−){
          Selected ← TCR_i;                                  TCR passes the thymus selection
          i++;
     fi
}
output(Selected = TCR_1, ..., TCR_N);                        TCRs that leave the thymus
```

Fig. 2. We randomly select $\mathtt{n_{self}}$ 9-mers that bind at least one of the $\mathtt{n_{mhc}}$ MHC molecules. Then we generate a random TCR and calculate the probability of being positively selected and those of being negatively selected. We iterate until we stochastically have N TCRs.

dependence is shown in figure 4. As expected, increasing the thymic efficiency E the average self reactivity α decreases. On the other hand for large values of E, α is less influenced by the number of self peptides $\mathtt{n_{self}}$.

The number of self peptides $\mathtt{n_{self}}$ therefore does not dramatically influence α but it does determine the probability for a T-cell to pass the selection instead. In other words it is related to the fraction of cells that leave the thymus. It has been estimated that each day of a young person 60×10^6 immature cells are tested but only 1 to 3% exit the thymus (Goldsby et al., 2000; Murphy et al., 2008). Figure 5 shows the ratio between the number of cells exiting the thymus ("out") and those entering the thymus ("in") against the number of self peptides $\mathtt{n_{self}}$. The figure shows that we can get reasonable out/in ratios with a number of self peptides between 100 and 200. Each point is the average of hundreds of independent runs with a randomly chosen set of self peptides. Note however that in general this curve depends on other parameters (E) and therefore is not too indicative of the real number $\mathtt{n_{self}}$ of self peptides.

An interesting question that has been already investigated long time ago with a computational model ((Celada & Seiden, 1992)) is to calculate the "optimal" number of MHC molecules. We performed similar experiments varying the number of MHC molecules and computing

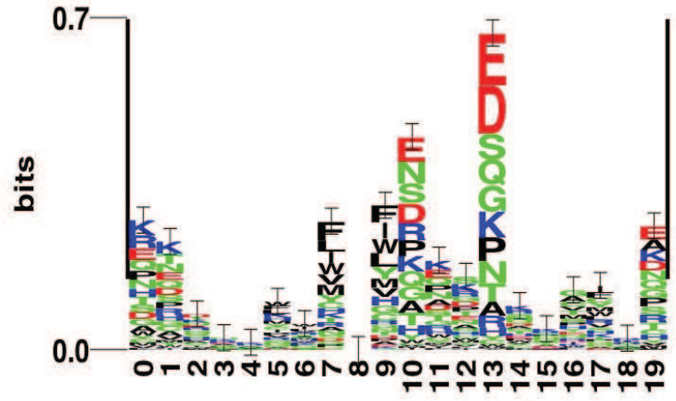

Fig. 3. Logo plots of the TRC sets that leave the thymus as operational CTLs TCRs. The height of the letters reflects the Shannon information at individual positions. Parameters $\mathtt{n_{self}} = 100$ and $\mathtt{E} = 10$. The MHC set is A*0201, A*6841, B*5304, B*5309. The Logo plots have been calculated using a small sample (i.e., 100) of the population of TCRs passing the selection.

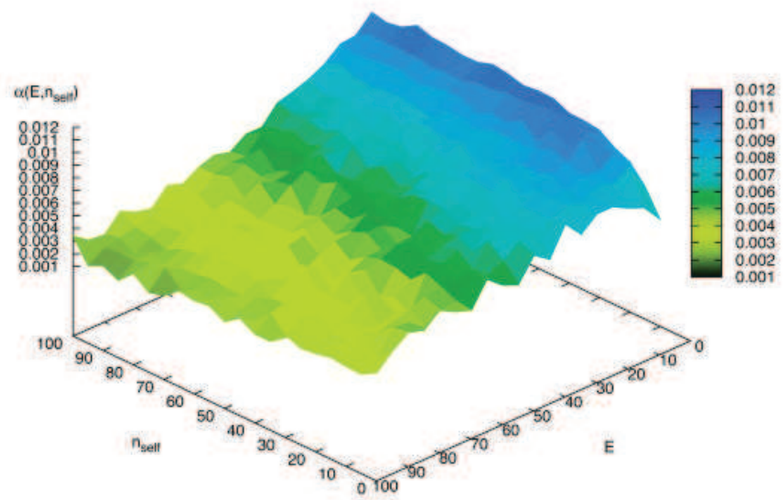

Fig. 4. Auto-reactivity rate $\alpha(\mathtt{E}, \mathtt{n_{self}})$ of the selected cells (those leaving the thymus) with respect to $\mathtt{E}$ and $\mathtt{n_{self}}$. Higher $\mathtt{E}$ corresponds to less average auto-reactivity. The influence of $\mathtt{n_{self}}$ is only marginal. The MHC alleles used are those reported in the caption of figure 3.

the average self reactivity $\alpha(\mathtt{n_{mhc}})$ of the selected TCRs. What we found is interesting though incorrect. In fact, the minimum of the self reactivity is attained for $\mathtt{n_{mhc}} = 8$ which is close to the real value of 6 alleles (figure 6). The overall curve is however in line with the ambivalent role of the MHCs: on the one hand, more MHCs foster the presentation of self peptides and, on the other hand, it limits the T-cell repertoire in the negative selection. An optimal number is therefore expected as the combination of the two opposed effects.

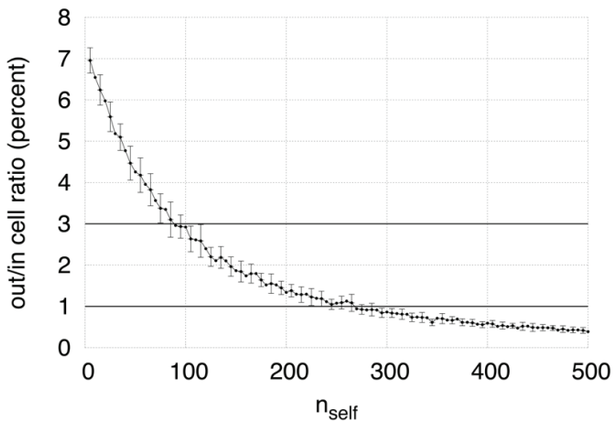

Fig. 5. The ratio between the number of cells entering the thymus ("in") and those leaving the thymus ("out") plotted against the number of self peptides $\mathtt{n_{self}}$. Parameter $\mathtt{E} = 10$. MHC-related parameters as in caption of figure 3. Parameters $\mathtt{E}$=60, $\mathtt{n_{self}}$=100.

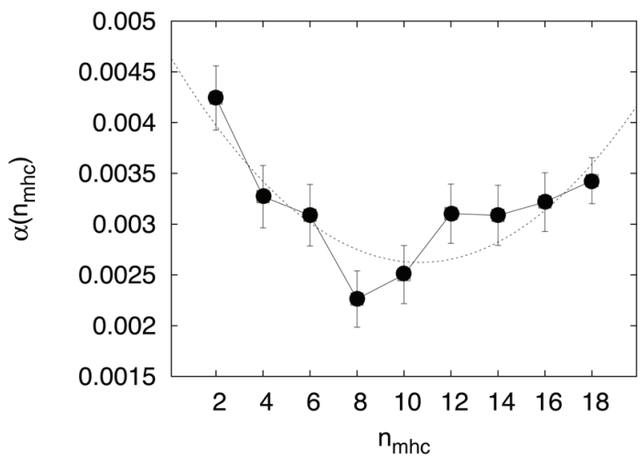

Fig. 6. This plot shows the average self-reactivity as a function of $\mathtt{n_{mhc}}$. Parameters $\mathtt{E}$=60, $\mathtt{n_{self}}$=100.

4. Conclusions

We have described a Monte Carlo simulation of the thymus where CTLs are selected on the basis of the affinity of their TCRs to MHC molecules and self peptides. We have used data-driven prediction tools to identify suitable self peptides for a specific MHC set of molecules. Furthermore, we have defined a general protein-protein binding potential on the basis of the work of Miyazawa and Jernigan on protein energy potentials (Miyazawa & Jernigan, 2000) that provides a method for assessing the chances of direct interactions among proteins.

Finally, the binding affinity between the TCR and the MHC-peptide complex is computed by taking into account a MHC class I specific contact matrix. This matrix was derived by finding

residues in contact between the α and β chains of the TCR and the HLA-A2 heavy chain of an MHC molecule in a high resolution 3D structure (access number 1OGA for Protein Data Bank).

By running a large number of simulations we have estimated the average auto-reactivity rate for a wide range of the parameters and found that auto-reactive cells are able to leave the thymus but their number and their overall ability to recognize self peptides decreases with $\mathtt{E}$, a parameter that indicates the "time" spent in the thymus. However this rate is only slightly affected by the number of self peptide presented $\mathtt{n_{self}}$.

Other simulation performed changing the number of MHC alleles $\mathtt{n_{mhc}}$ resulted in a minimum of average self-reactivity for $\mathtt{n_{mhc}} = 8$ which is wrong but not too distant from reality.

Take together, these results show that the simulation performs reasonably well. This is encouraging given the number of working assumptions that we had to make at this stage.

This study follows the lines of Morpurgo *et al.* ((Morpurgo et al., 1995)) but diverges from it in that the molecules represented are not binary strings but rather sequences of amino acids and, most importantly, the function used to compute the affinity among these molecules is provided by data-driven machine learning bioinformatics methods. We have already adopted this approach in a previous (Rapin et al., 2010); we believe that, although preliminary and somehow approximate at this stage, it provides a promising way to incorporate immuno-informatics resources (both data and methods) to systemic level stochastic simulations of immunological processes.

5. References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* **25**(17): 3389–402.

Arstila, T., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J. & Kourilsky, P. (1999). A direct estimate of the human t cell receptor diversity, *Science* 286(5441): 958–961.

Buus, S., Lauemoller, S. L., Worning, P., Kesmir, C., Frimurer, T., Corbet, S., Fomsgaard, A., Hilden, J., Holm, A. & Brunak, S. (2003). Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach, *Tissue Antigens* 62(5): 378–384.

Celada, F. & Seiden, P. E. (1992). A computer model of cellular interactions in the immune system., *Immunol Today* 13(2): 56–62.

Darve, E. & Pohorille, A. (2001). Calculating free energies using average force, *J Chem Phys* 115: 9169–9183.

Efroni, S., Harel, D. & Cohen, I. (2005). Reactive animation: Realistic modeling of complex dynamic systems, *Computer* 38: 38–47.

Efroni, S., Harel, D. & Cohen, I. (2007). Emergent dynamics of thymocyte development and lineage determination, *PLoS Computational Biology* 3: 127–135.

Efroni, S., Harel, D. & Cohen, I. R. (2003). Toward rigorous comprehension of biological complexity: modeling, execution, and visualization of thymic t-cell maturation., *Genome Res* 13(11): 2485–2497.

Germain, R. (2002). T-cell development and the cd4-cd8 lineage decision, *Nat Rev Immunology* 2: 309–322.

Goldsby, R., Kindt, T. & Osborne, B. (2000). Kuby immunology, iv ed.

Hazenberg, M., Otto, S., Stuart, J., Verschuren, M., Borleffs, J., Boucher, C., Coutinho, R., Lange, J., de Wit, T., Tsegaye, A., van Dongen, J., Hamann, D., de Boer, R. & Miedema, F.

(2000). Increased cell division but not thymic dysfunction rapidly affects the t- cell receptor excision circle content of the naive t cell population in hiv-1 infection, *Nat Medicine* 6: 1036–1042.

Klein, L., Hinterberger, M., Wirnsberger, G. & Kyewski, B. (2009). Antigen presentation in the thymus for positive selection and central tolerance induction, *Nat Rev Immunol* 9: 833–844.

Korber, B., LaBute, M. & Yusim, K. (2006). Immunoinformatics comes of age, *PLoS Comp Biol* 2(6): 484–492.

Lin, H. H., Ray, S., Tongchusak, S., Reinherz, E. L. & Brusic, V. (2008). Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research, *BMC Immunol* 9: 8.

Lin, H. H., Zhang, G. L., Tongchusak, S., Reinherz, E. L. & Brusic, V. (2008). Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research, *BMC Bioinformatics* 9 Suppl 12: S22.

Lund, O., Kesmir, C., Nielsen, M., Lundegaard, C. & Brunak, S. (2005). *Immunological Bioinformatics*, MIT Press, Cambridge, Mass.

Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., Buus, S. & Brunak, S. (2004). Definition of supertypes for HLA molecules using clustering of specificity matrices., *Immunogenetics* 55(12): 797–810.

Lundegaard, C., Lund, O., Kesmir, C., Brunak, S. & Nielsen, M. (2007). Modeling the adaptive immune system: predictions and simulations., *Bioinformatics* 23(24): 3265–3275.

Mehr, R., Abel, L., Ubezio, P., Globerson, A. & Agur, Z. (1993). A mathematical model of the effect of aging on bone marrow cells colonizing the thymus, *Mechanisms of Aging and Development* 67: 159–172.

Mehr, R., Globerson, A. & Perelson, A. (1995). Modeling positive and negative selection and differentiation processes in the thymus, *J Theor Biol* 175: 103–126.

Mehr, R., Perelson, A., Fridkis-Hareli, M. & Globerson, A. (1997). Regulatory feedback pathways in the thymus, *Immunol Today* 18: 581–585.

Mehr, R., Perelson, A., Sharp, A., Segel, L. & Globerson, A. (1998). Mhc-linked syngeneic developmental preference in thymic lobes colonized with bone marrow cells: A mathematical model, *Dev. Immunol* 5: 303–318.

Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading., *J Mol Biol* 256(3): 623–644.

Miyazawa, S. & Jernigan, R. L. (1999). An empirical energy potential with a reference state for protein fold and sequence recognition., *Proteins* 36(3): 357–369.

Miyazawa, S. & Jernigan, R. L. (2000). Identifying sequence-structure pairs undetected by sequence alignments, *Protein Eng.* 13(7): 459–475.
URL: *http://peds.oxfordjournals.org/cgi/content/abstract/13/7/459*

Morpurgo, D., Serenthà, R., Seiden, P. E. & Celada, F. (1995). Modelling thymic functions in a cellular automaton., *Int Immunol* 7(4): 505–516.

Murphy, K., Travers, P., Janeway, C. & Mark, W. (2008). *Janeway's Immunology*, Garland Science, Taylor and Francis, New York.

Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Roder, G., Peters, B., Sette, A., Lund, O. & Buus, S. (2007). NetMHCpan, a method for

quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence., *PLoS ONE* 2(8): e796.

Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S. & Lund, O. (2004). Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach., *Bioinformatics* 20(9): 1388–1397.

Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S., Lamberth, K., Buus, S., Brunak, S. & Lund, O. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations, *Protein Sci* 12(5): 2007–2017.

Nishio, J., Suzuki, M., Nanki, T., Miyasaka, N. & Kohsaka, H. (2004). Development of tcrb cdr3 length repertoire of human t lymphocytes, *Int Immunol* 16(3): 423–431.

Petrovsky, N. & Brusic, V. (2002). Computational immunology: The coming of age, *Immunol Cell Biol* 80: 248–254.

Petrovsky, N. & Brusic, V. (2006). Bioinformatics for study of autoimmunity, *Autoimmunity* 39: 635–643.

Rapin, N., Lund, O., Bernaschi, M. & Castiglione, F. (2010). Computational immunology meets bioinformatics: the use of prediction tools for molecular binding in the simulation of the immune system, *PLoS ONE* 5(4): e9862. doi:10.1371/journal.pone.0009862.

Souza-e Silva, H., Savino, W., Feijóo, R. & Vasconcelos, A. (2009). A cellular automata-based mathematical model for thymocyte development, *PLoS ONE* 4(12): e8233. doi:10.1371/journal.pone.0008233.

von Boehmer, H. & Kisielow, P. (1993). Lymphocyte lineage commitment: Instruction versus selection, *Cell* 73: 207–208.

Yewdell, J., Reits, E. & Neefjes, J. (2003). Making sense of mass destruction: quantitating mhc class i antigen presentation, *Nat Rev Immunol* 3(12): 952–961.

Yewdell, W, J. & Bennink, J. R. (1999). Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses., *Annu Rev Immunol* 17: 51–88.

Applications of Monte Carlo Methods in Biology, Medicine and Other Fields of Science

Edited by Prof. Charles J. Mode

ISBN 978-953-307-427-6
Hard cover, 424 pages
Publisher InTech
Published online 28, February, 2011
Published in print edition February, 2011

This volume is an eclectic mix of applications of Monte Carlo methods in many fields of research should not be surprising, because of the ubiquitous use of these methods in many fields of human endeavor. In an attempt to focus attention on a manageable set of applications, the main thrust of this book is to emphasize applications of Monte Carlo simulation methods in biology and medicine.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Filippo Castiglione (2011). A Monte Carlo Simulation for the Construction of Cytotoxic T Lymphocytes Repertoire, Applications of Monte Carlo Methods in Biology, Medicine and Other Fields of Science, Prof. Charles J. Mode (Ed.), ISBN: 978-953-307-427-6, InTech, Available from: http://www.intechopen.com/books/applications-of-monte-carlo-methods-in-biology-medicine-and-other-fields-of-science/a-monte-carlo-simulation-for-the-construction-of-cytotoxic-t-lymphocytes-repertoire

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen