Published in final edited form as:

Theor Popul Biol. 2011 December ; 80(4): 233–243. doi:10.1016/j.tpb.2011.08.005.

Optimal viral immune surveillance evasion strategies

Alexandra Agranovich¹, Tal Vider-Shalit¹, and Yoram Louzoun^{1,2}

¹Department of Mathematics and Gonda Brain Research Center, Bar-Ilan University, Ramat Gan 52900, Israel

Abstract

Following cell entry, viruses can be detected by Cytotoxic T Lymphocytes. These Cytotoxic T Lymphocytes can induce host cell apoptosis and prevent the propagation of the virus. Viruses with fewer epitopes have a higher survival probability, and are selected through evolution. However, mutations have a fitness cost and on evolutionary periods viruses maintain some epitopes. The number of epitopes in each viral protein is a balance between the selective advantage of having less epitopes and the reduced fitness following the epitope removing mutations. We discuss a bioinformatic analysis of the number of epitopes in various viral proteins and propose an optimization framework to explain these numbers. We show, using a genomic analysis and a theoretical optimization framework, that a critical factor affecting the number of presented epitopes is the expression stage in the viral life cycle of the gene coding for the protein. The early expression of epitopes can lead to the destruction of the host cell before budding can take place. We show that a lower number of epitopes is expected in early proteins even if late proteins have a much higher copy number.

Keywords

virus; epitope; SIR score; optimization

1 INTRODUCTION

Viral infections can be countered by two arms of the immune system. The humoral immune system removes free virions using antibodies. The cellular immune system removes infected cells (Harty et al., 2000; Metchnikoff and Binnie, 1905). The main effect of the cellular immune system is mediated by Cytotoxic T Lymphocytes (CTL), which can induce the apoptosis of infected cells. CTLs can recognize viral epitopes only in the context of MHC-I molecules (Sebzda et al., 1999; van den Broek and Hengartner, 2000). Viruses use the cellular protein transcription mechanisms to produce their own proteins in the cytosol. These proteins are then degraded by proteasomes into short peptides (Rock *et al.*, 2002). The cleaved peptides can bind to the energy dependent TAP channels embedded in the ER membrane, transfer to the ER cavity and bind to the alpha chain of the MHC-I complex. The peptide-MHC complex is then transferred through the Golgi apparatus to the plasma membrane (Braciale and Braciale, 1991; Kleijmeer et al., 1992; Townsend and Bodmer, 1989). The recognition of viral epitopes by CTLs leads to the host cell apoptosis (Janeway,

© 2011 Elsevier Inc. All rights reserved.

²Correspondence and requests should be addressed to Yoram Louzoun. louzouy@math.biu.ac.il, phone-972-3-5317610.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

2001). Thus, if such recognition occurs before viral budding (or bursting, depending on the virus type), the cell hosting the virus will be destroyed before virions can infect new cells.

Viruses have evolved a wide spectrum of techniques to evade the immune system detection. One of them is the accumulation of mutations that remove T-cell epitopes (Aebischer et al., 1991; Bowen and Walker, 2005; McMichael and Phillips, 1997). The MHC-I molecule has clear preferences for the amino acid composition of binding peptides. Mutations in these positions can significantly reduce the MHC-peptide binding affinity and prevent peptide presentation. Mutation mediated viral immune surveillance evasion was first acknowledged by the study of the murine LCMV infection (Pircher *et al.*, 1990). Such mutations do not come free of cost. Mutations leading to epitope removal induce changes in the amino acid composition that can lead to a lower viral fitness (Rimmelzwaan *et al.*, 2005; Soderholm *et al.*, 2006). Thus a balance exists between the viral survival advantage following random epitope removal and the decrease in the fitness induced by the same mutations. In this paper, we propose a mathematical model to study the effects of this balance on the epitope number distribution among viral proteins, assuming the studied virus has reached its evolutionary optimal steady state. To the best of our knowledge, this is the first such theoretical analysis. The proposed model takes into account the expression time and copy number of each viral protein, as well as the CTL dynamics.

We have previously performed a bioinformatic analysis of the epitope density in viral proteins (Vider-Shalit et al., 2007b; Vider-Shalit et al., 2009a; Vider-Shalit et al., 2009b). We found that the epitope density is non-uniform among proteins of the same virus. Specifically, we have used a combination of bioinformatic algorithms to compute all presented epitopes in a large group of viral proteins in multiple HLA alleles. Presented epitopes were computed as ninemers predicted to be cleaved, bind to TAP and bind to a given MHC allele. For details of the precise algorithms, see Appendix A. We then computed the epitope density in different proteins (i.e. the number of epitopes divided by the number of possible ninemers in the protein is called *SIR* score). We compared early to late proteins in 24 different viruses, including Adenoviruses, Herpes Viruses, HIV and HPV. Late and early proteins are defined based on the presence of early and late expression gene cassettes. In most HLA alleles that were tested, the average *SIR* score ratio of late to early proteins is higher than 1. This result is valid for most viruses, as well as for the average of all viruses (Fig. 1, adapted from (Vider-Shalit, 2009c)). The difference between early and late proteins is significant, when comparing the average epitope density over all HLA alleles simultaneously (T test, $p \leq 0.0001$), or when the *SIR* score for each allele is taken into account (T test, $p < 1.e-100$). Thus, quite systematically, late viral proteins express more epitopes than early ones. Note that not all peptides presented on MHC molecules can actually activate T cells. Moreover, not all T cell responses actually clear the virus. Thus the presence or absence of epitopes is not a direct proof for the destruction of the infected population. It is however, indirect evidence that the virus is affected by the immune response against certain proteins. A clear example is the case of EBNA that induces an immunodominant response (Steven et al., 1997), but still does not lead to the clearance of the latent EBV.

Following the bioinformatic analysis, we here propose a theoretical population dynamics framework to study the effect of different factors in the viral life-cycle on the expected epitope distribution.

2 THE MODEL

We model a virus as a set of different proteins expressed in a host cell. Each protein i has its own expression profile: $x_i(t)$ is the concentration of protein i at time t . Each such protein can

induce a polyclonal T cell response $T_i(t)$. Each T_i represents the number of cells in a group of clones reacting to the epitopes in a given protein. For the sake of simplicity we denote T_i as a clone, although it is not a single clone in the strict sense of the word, since it is composed of multiple clones reacting to possibly different epitopes on the same protein. Finally, we denote by p_i the number of *different* epitopes presented by each viral protein. At the current stage, we treat all epitopes as equal; we will relax this assumption later in the manuscript.

The viral – T lymphocytes dynamics can be simplified by the following multidimensional ODE:

$$\begin{cases} \frac{d\mathbf{x}}{dt} = h(\mathbf{x}) \\ \frac{dT}{dt} = f(\mathbf{p}, \{\mathbf{x}\}, T) \end{cases} \quad (1)$$

where $\mathbf{x}$, $\mathbf{p}$ and T are the vectors containing $x_i(t)$, p_i and T_i . $h(\mathbf{x})$ represents the dynamics of the viral proteins following cell entry in a given cell, and $f(\mathbf{p}, \{\mathbf{x}\}, T)$ represents the dynamics of T cell clones during the course of a disease. Note that T and $\mathbf{x}$ evolve on very different time scales (Fig. 2). The dynamics of $\mathbf{x}$ is the dynamics of proteins inside an infected cell with a typical time scale of hours. These dynamics end at cell death or following budding. The dynamics of T cells evolve at a much slower time scale, and span many viral life cycles. Moreover, the dynamics of T are affected by all infected cells. We denote the set of the $x_i(t)$ values in all infected cells as $\{\mathbf{x}\}$. To summarize, we assume the following simplistic model of the viral population dynamics:

- A virus infects a new host cell.
- The host cell expresses viral proteins, each with a different expression pattern.
- After a given time ($t_{budding}$), the full virions leave the cell, and the host cell dies.
- The host cell can die before $t_{budding}$ because it is killed by immune system cells, or the viruses may fail to integrate. In both cases no viral progeny is produced.

Within this model, the basic question we address is what is the optimal epitope distribution required to maximize the probability that a virus infecting a cell will have a progeny.

The virus host dynamics can be treated as an epidemic model, where each cell is susceptible. The virus can spread among the host cells, if each infected cell infects on average at least one susceptible cell. This is the parallel of having an R_0 value of more than 1 in the SIR model (Anderson and May, 1991).

We here test the effect of two aspects of viral protein dynamics on the optimal epitope number: the expression time and the total protein copy number. We assume a simplistic model of viral protein dynamics to test these effects. Based on our initial results, we expect the effect of the expression time to be more crucial. In order to assess the host cell death probability, the T cell clones size dynamics should be considered, as will be further discussed.

If a given protein has x_i copies and p_i epitopes per copy, it presents a total of $p_i x_i$ epitopes. One can roughly estimate the total number of epitopes presented by a virus in a cell as

$\sum_i p_i x_i(t)$, assuming no large differences in the affinity of each peptide. Further assuming a constant killing rate per T cell and per epitope (μ_0), the total killing rate before budding of a given virally infected cell can be estimated by:

$$\mu(\mathbf{p}) = \mu_0 \int_0^{t_{budding}} dt \sum_i p_i x_i T_i. \quad (2)$$

The probability that this host cell would survive for a long enough time allowing viral budding is thus:

$$P(\text{survival}) = \exp \left(-\mu_0 \int_0^{t_{budding}} dt \sum_i p_i x_i(t) T_i(t) \right). \quad (3)$$

One can assume that on evolutionary time scales, viruses have evolved to maximize this survival probability. We propose to compute the distribution of $\mathbf{p}$ maximizing $P(\text{survival})$. Other factors obviously affect the survival probability, but they do not depend on $\mathbf{p}$ and will not affect the optimization results.

The simplest distribution of $\mathbf{p}$ would obviously be a constant value of zero epitopes for all viral proteins. However, given the fitness cost of adaptations, a more complex picture emerges. Assuming each mutation has a multiplicative cost on the survival probability, the survival probability is:

$$P(\text{survival}) = \prod_i d^{p_{i0} - p_i} \exp \left(-\mu_0 \int_0^{t_{budding}} dt \sum_i p_i x_i T_i \right) \quad (4)$$

where d represents the contribution of each mutation to the death probability ($d < 1$) and p_{i0} is the basal number of epitopes of protein i . The results of the optimization will be similar even if a more complex function of the mutations is assumed:

$P(\text{survival}) = \phi \left(\sum_i p_i \right) \cdot \exp \left(-\mu_0 \int_0^{t_{budding}} dt \sum_i p_i x_i T_i \right)$, as long as ϕ is a positive function with a minimum at 0. We do not consider here the possibility of adaptations increasing the number of different epitopes. Instead of maximizing $P(\text{survival})$, $-\log(P(\text{survival}))$ is minimized, leading to the following minimization problem:

$$L(p^*) = \min_{p_i \geq 0} \left(\mu_0 \int_0^{t_{budding}} dt \sum_i p_i x_i T_i + \bar{d} \sum_i (p_{i0} - p_i) \right), \quad (5)$$

where $\bar{d} = -\ln d > 0$.

This problem is formally similar to assuming a constraint on the number of epitopes

$$\left(\sum_i p_i > c \right) \text{ and minimizing } \mu_0 \int_0^{t_{budding}} dt \sum_i p_i x_i T_i.$$

The epitope density distribution maximizing the probability that a virus will produce a progeny following a viral life cycle is affected by the T cell distribution. We present two extreme cases. The first case occurs when a large number of cells are infected, and the protein expression in a given cell has no effect on the T cell population. The second case is when very few cells are infected and the T cells clones are small. In such a case, the T cell

clone growth rate is determined by the activation of T cells, which is in turn determined by the epitopes presented.

In the first case, the viral entry time is different for each infected cell, and the expression time of viral proteins is not synchronized among cells (Fig. 2). We denote this case as the asynchronous model. The second case represents the initial stages of the disease, where we assume a small number of virions seeding host cells. We assume that in this case, all cells in a given generation of infected cells are infected simultaneously. This case is denoted as the synchronous model.

To summarize, beyond the viral dynamics, we assume the following cell host killing process

- Infected host cells can activate CTLs.
- CTLs can detect epitopes on infected host cells.
- The detection probability is proportional to the product of the number of epitopes in each protein and the concentration of the protein carrying the epitopes.
- The destruction probability is proportional to the product of the T cells clone size and the detection probability.

3 GENERAL SOLUTION

The minimization problem (5) can be translated to a positive constrained minimization problem:

$$\mu(\mathbf{p}^*) = \min_{\sum_i p_i \geq c, \mathbf{p} \geq \mathbf{0}} \mu(\mathbf{p}). \quad (6)$$

The first constraint in Eq. (6) represents the fitness cost of adaptation, and c represent the minimal number of epitopes that must be maintained. A constraint on c is equivalent to a constraint on the total number of epitope removing mutations.

In the synchronous model that represents the earliest phases of the infection, we assume that the T cells are being activated by the epitopes and that their population is low enough to ignore possible saturation terms. The asynchronous model represents the late or chronic stages of the disease, where we assume that the T cell populations are in steady state. In such a case, their concentration is determined by the average activation signal they receive. These considerations lead to the following simplified dynamics of two concerned cases:

$$\begin{cases} \frac{dx}{dt} = h(\mathbf{x}) \\ \frac{dT_i}{dt} = p_i x_i T_i, \quad 1 \leq i \leq n \end{cases} \quad \text{or} \quad \begin{cases} \frac{dx}{dt} = h(\mathbf{x}) \\ T_i = \langle p_i x_i \rangle, \quad 1 \leq i \leq n \end{cases} \quad (7)$$

Thus, the exponential growth model describes the synchronous case. The second model describes the case where the T cell populations have reached a steady state and is denoted as the asynchronous model.

In all our models, we explicitly assume that all virions bud before some particular time $t_{budding}$. We show in Appendix E the results of an opposite model where the cell lives for a long time and keep budding virions. In such a model, the results are much more sensitive to the total protein copy number than the expression timing.

3.1 Asynchronous model

A long time after the initial infection, the population size of each T cell clone can be seen as static through the life cycle of a virus (Fig. 2) and proportional to the T cells average activation rate. This activation rate is in turn proportional to the average copy number of the appropriate viral protein. The system dynamics can be described as:

$$\begin{cases} \frac{dx}{dt} = h(x) \\ T_i^{as} = T^{as} p_i \int_0^{t_{budding}} x_i dt, \quad 1 \leq i \leq n \end{cases} \quad (8)$$

where we name the constant concentration of T lymphocytes of type i as T_i^{as} . $\mu(\mathbf{p})$ can be

computed by substituting the solution of Eq. (8) into Eq. (6) to get $\mu(\mathbf{p}) = C_1 \cdot \sum_i A_i \cdot (p_i)^2$,

where $A_i = g_i^2$ and $g_i = \int_0^{t_{budding}} x_i dt$. We look for the set $\mathbf{p}^*$ satisfying

$$\mu(\mathbf{p}^*) = \min_{\sum_i p_i \geq c, \mathbf{p} \geq \mathbf{0}} \sum_i A_i \cdot p_i^2 \quad (9)$$

Solution of optimization problem—For all i , $A_i > 0$ and the optimization problem is a constrained convex quadratic optimization problem that can be solved using the KKT (Karush-Kuhn-Tucker) theorem (Karush, 1939; Kuhn and Tucker, 1951), leading to a unique positive minimum (Luenberger and Ye, 2008).

The inequality constraint $\sum_i p_i \geq c$ is always active ($\sum_i p_i = c$). We can thus write

$p_{i_0} = c - \sum_{i \neq i_0} p_i$ where i_0 is the index of some nonzero element p_i , and reformulate the optimization problem (9):

$$\min_{\mathbf{p} \geq \mathbf{0}} \left(\sum_{i \neq i_0} A_i \cdot (p_i)^2 + A_{i_0} \cdot \left(c - \sum_{i \neq i_0} p_i \right)^2 \right). \quad (10)$$

The Lagrangian is:

$$L(\mathbf{p}, \eta) = \sum_{i \neq i_0} A_i \cdot (p_i)^2 + A_{i_0} \cdot \left(c - \sum_{i \neq i_0} p_i \right)^2 - \sum_{i \neq i_0} \eta_i p_i. \quad (11)$$

Up to a constant difference, this is equivalent to the maximization problem in Eq.(5). The KKT equations are:

$$\frac{\partial L}{\partial p_i} = 2A_i p_i - 2A_{i_0} \left(c - \sum_{i \neq i_0} p_i \right) - \eta_i = 0, \quad 1 \leq i \leq n, i \neq i_0, \quad (11a)$$

$$\frac{\partial L}{\partial \eta_i} = -p_i \leq 0, \quad 1 \leq i \leq n \wedge i \neq i_0, \quad (11b)$$

$$\eta_i p_i = 0, \quad 1 \leq i \leq n \wedge i \neq i_0, \quad (11c)$$

$$\eta_i \geq 0, \quad 1 \leq i \leq n \wedge i \neq i_0. \quad (11d)$$

From (11a) we get $\eta_i = 2A_i p_i - 2A_{i_0} \left(c - \sum_{i \neq i_0} p_i \right)$. The strictly positive solution can be obtained if $\eta_i = 0$ for all $1 \leq i \leq n \wedge i \neq i_0$, yielding $p_i = \frac{A_{i_0}}{A_i} p_{i_0}$, $1 \leq i \leq n \wedge i \neq i_0$ that when substituted into condition $\sum_{i=1}^n p_i = c$, takes the form

$$p_i = \frac{c/z}{A_i}, \quad (12)$$

where $z = \sum_{j=1}^n \frac{1}{A_j}$ (i.e. the number of epitopes is inversely proportional to A_i). We prove in Lemma 1 (see Appendix B) that in the optimum solution, all p_i are strictly positive. We will further discuss the numerical results obtained for various models of the viral protein dynamics $h(x)$.

3.2 Synchronous model

In some viral infections, such as HIV, long term reservoirs of infected cells are produced at the earliest stages of the infection (Haase, 1999; Rehmann and Nascimbeni, 2005; van Rompay et al., 1999). Thus, this very early stage may be the most critical from the viral perspective. In the initial period of the infection, the T cell clone size growth can be described by a mass-action formalism (Guldberg, 1899; Sykulev *et al.*, 1995). Furthermore, assuming the virions simultaneously invade a very limited number of host cells, we compute the T cell dynamics with all viral proteins synchronized to obtain:

$$\begin{cases} \frac{dx}{dt} = h(x) \\ \frac{dT_i}{dt} = p_i x_i T_i, \quad 1 \leq i \leq n \end{cases} \quad (13)$$

Eq. (13) can be substituted into (6) to produce the non-linear convex optimization problem:

$$\min_{p \geq 0, \sum_{i=1..n} p_i \geq c} (\mu) = \min_{i=1..n} \sum T_i^0 A_i^{p_i}; A_i = e^{g_i}; g_i = \int_0^{t_{budding}} x_i dt. \quad (14)$$

Solution of optimization problem—Eq. (14) is a non-linear optimization problem and can be solved again using the KKT (Karush-Kuhn-Tucker) conditions (Karush, 1939; Kuhn and Tucker, 1951). Since it is a convex optimization problem, the KKT point is the unique minimum (Luenberger and Ye, 2008). The Lagrangian of Eq. (14) is

$$L(\mathbf{p}, \eta, \nu) = \sum_i T_i^0 A_i^{p_i} - \nu \left(\sum_i p_i - c \right) - \sum_i \eta_i p_i. \quad (15)$$

For the entire development of KKT see Appendix C. From KKT we get the following solution for the active constraints:

$$\eta_i = T_i^0 A_i^{p_i} \ln A_i - T_{i_0}^0 A_{i_0}^{c - \sum_{i \neq i_0} p_i} \ln A_{i_0}, \quad 1 \leq i \leq n \wedge i \neq i_0, \quad (16)$$

where i_0 is the index of a protein with a positive number of epitopes (There must be at least one such protein). The fraction of active constraints is a function of T_i^0 and A_i . We have developed a rapid algorithm to solve the optimization problem in this case, without checking the 2^n possibilities induced by the KKT solutions (see Appendix D). For all non-zero solutions, the value of p_i is:

$$p_i = \ln \left(\frac{T_{i_0}^0 \ln A_{i_0}}{T_i^0 \ln A_i} A_{i_0}^{p_{i_0}} \right) / \ln A_i. \quad (17)$$

Eq. (17) can be simplified by defining $Z_i = \frac{T_{i_0}^0 \ln A_{i_0}}{T_i^0 \ln A_i}$ to obtain:

$$p_i = \frac{1}{\ln A_i} \left[\ln Z_i + \left(c - \sum_j \frac{\ln Z_j}{\ln A_j} \right) / \left(\sum_j \frac{1}{\ln A_j} \right) \right] = \frac{\ln Z_i + C_1}{\ln A_i}, \quad (18)$$

where C_1 is a constant.

3.3 Comparison between the synchronous and the asynchronous models

A few basic differences exist between the solutions of the two models. Both models are fully

determined by the value of $g_i = \int_0^{t_{budding}} x_i dt$. Thus the optimal epitope distribution is not a function of the peak density of each protein or of its precise expression pattern. The only thing that affects the distribution is the integral of the copy number until the virus can bud.

However, one can already see from the general solution that the synchronous model has a sharper distribution, with some proteins having no epitopes and other having all the epitopes. In contrast in the asynchronous models, all p_i are strictly positive.

Note that in both models, either increasing the total copy number or decreasing the initial expression time will increase g_i and lead to a lower optimal epitope number as is indeed observed. An interesting question remaining is the effect of changing both factors simultaneously in opposite directions. We here checked a couple of simplified models for

the protein dynamics to study these effects. Given more realistic protein expression patterns, one can simply compute g_i for each protein and assess the optimal epitope distribution. A basic conclusion from the model up to now is however that given two proteins with different values of g_i , the one with the higher g_i is expected to have less epitopes.

4 SIMPLIFIED PROTEIN DYNAMICS

In order to compare the effect of the expression time and the protein copy number, we used viral protein dynamics containing only these two elements. We studied two very simple (and obviously not realistic) models for the viral proteins: logistic growth (Verhulst, 1838) and a constant expression rate with self destruction. These two models are characterized by a expression rate and a saturation value, allowing us to study the balance between these two effects.

In the viral protein logistic growth model, the copy number expression rate is defined as λ_i and the saturation value as λ_i/σ_i ; in the constant growth model, the expression rate is defined as β_i and the decay rate as δ_i , leading to the following dynamics:

$$\frac{dx_i}{dt} = \lambda_i x_i \left(1 - \frac{\sigma_i}{\lambda_i} x_i\right) \text{ or } \frac{dx_i}{dt} = \beta_i \left(1 - \frac{\delta_i}{\beta_i} x_i\right). \quad (19)$$

For the asynchronous model, we obtain for the logistic growth:

$$A_i = \frac{1}{\sigma_i^2} \ln^2 \left(1 + \frac{\sigma_i}{\lambda_i} x_i^0 (e^{\lambda_i t_{budding}} - 1)\right). \quad (20)$$

and for the constant growth:

$$A_i = \left(\frac{\beta_i}{\delta_i} t_{budding} + \left(\frac{\beta_i}{\delta_i^2} - \frac{x_i^0}{\delta_i} \right) (e^{-\delta_i t_{budding}} - 1) \right)^2. \quad (21)$$

For the synchronous model, we obtain for the logistic growth:

$$A_i = \left(1 + \frac{\sigma_i}{\lambda_i} x_i^0 (e^{\lambda_i t_{budding}} - 1)\right)^{1/\sigma_i}. \quad (22)$$

and for the constant growth:

$$A_i = \exp \left(\frac{\beta_i}{\delta_i} t_{budding} + \left(\frac{\beta_i}{\delta_i^2} - \frac{x_i^0}{\delta_i} \right) (e^{-\delta_i t_{budding}} - 1) \right). \quad (23)$$

These expressions can be used to compute the optimal distribution in equations (9) and (14). We here detail the results for the logistic growth model. The other model leads to similar conclusions.

We assume a low initial value for each protein ($x_i^0=0.1$) and that budding occurs after the expression of the last protein. If the saturation level is constant (we set $\lambda_i/\sigma_i = 1$), A_i is an

increasing function of λ_i and p_i is a decreasing function of λ_i (Fig. 3A). The more interesting situation occurs when early expressed proteins have a lower protein copy number. In such a case, the low density could have been expected to balance the risk induced by the early exposure. A simple example would be to set $\lambda_i/\sigma_i \propto 1/\lambda_i$; in other words, setting the saturation level to be inversely proportional to the expression level. In such a case, early expressed proteins would have a low saturation level. However, even in this model early expressed proteins have less epitopes in the optimal solution both in the asynchronous (Fig. 3A) and the synchronous case (Fig. 4A). The results for the simple growth model are very similar to the one observed in the logistic growth model (Fig. 3B and Fig. 4B).

5 EFFECT OF MULTIPLE ROUNDS OF CELL INFECTION

After the initial set of cells is infected, these cells can produce virions, which will in turn infect other cells. We have tested if the conclusions on the epitope distribution in the first generation hold in the following generations of infected cells. We here simplify the dynamics by assuming that these infection rounds are discrete. This simplification obviously fails after a few generations when infection events become unsynchronized, but may be true for a limited number of infection rounds. We thus solve a model for the firsts rounds of infection (we denote each such round as a generation).

We assume a uniform initial concentration of all CTL types in the first generation among all proteins. In the following generations, the T cell concentration at the beginning of the generation is affected by the previous generations and is thus non-equal among proteins. Using Eq. (13) and (22), the concentration of CTL type i at generation m can be computed as follows:

$$T_i = T_i^0 A_i^{(m-1)p_i} \cdot a_i^{p_i}(t), \quad 1 \leq i \leq n, \quad (24)$$

where $x^0 = x(0)$, $T^0 = T(0)$ are the initial conditions and the function $a_i(t)$ has the same expression as constant A_i where $t_{budding}$ is replaced by t . Eq. (24) can be substituted into (6) to produce the non-linear convex optimization problem:

$$\min_{p \geq 0, \sum_{i=1..n} p_i \geq c} (\mu) = \min \left(\sum_{i=1..n} T_i^0 \left(A_i^{mp_i} - A_i^{(m-1)p_i} \right) \right), \quad (25)$$

where $A_i = \left(\frac{\sigma_i x_i^0}{\lambda_i} e^{\lambda_i t_{budding}} - \frac{\sigma_i x_i^0}{\lambda_i} + 1 \right)^{\frac{1}{\sigma_i}}$ in the case of the logistic protein growth and $A_i = \exp \left(\frac{\beta_i}{\delta_i} t_{budding} + \left(\frac{\beta_i}{\delta_i^2} - \frac{x_i^0}{\delta_i} \right) (e^{-\delta_i t_{budding}} - 1) \right)$ in the case of the constant protein source.

The optimization problem is solved numerically and the results are weakly affected by the generation number (Fig. 5).

6 EFFECT OF EPITOPE AFFINITY

A major assumption in the entire optimization analysis was that of equal affinity of all epitopes. If this assumption is significantly relaxed, the difference between proteins decreases. An extreme model of epitope differences is immunodominance, where we assume that a single epitope induces a T cell response in each protein. In order to model differences

between epitopes, we assume n viral proteins and m different epitopes for each protein. We denote by p_{ij} the probability that epitope j of protein i is presented. The expected number of epitopes in protein i is simply:

$$p_i = \sum_{j=1}^m p_{ij}; 1 \leq i \leq n. \quad (26)$$

Equation (13) becomes in the synchronous model:

$$\begin{cases} \frac{dx_i}{dt} = h(x) \\ \frac{dT_i}{dt} = x_i T_i \sum_{j=1}^m \varepsilon^j p_{ij}, \quad \varepsilon \geq 1; 1 \leq i \leq n \end{cases} \quad (27)$$

For the sake of simplicity, we assume a geometric decrease in epitope presentation with a ratio ε (denoted as the *immunodominance factor*).

The optimization problem becomes

$$\min_{\substack{\sum_{i,j} p_{ij} \geq c, 0 \leq p_{ij} \leq 1}} \mu(t) = \int_0^{t_{budding}} dt' \sum_i x_i(t') T_i(t') \varepsilon^j p_{ij} = \sum_i T_i^0 \cdot A_i^{\sum_j \varepsilon^j p_{ij}} \quad (28)$$

where $A_i = e^{g_i}$; $g_i = \int_0^{t_{budding}} x_i dt$ or specifically $A_i = \left(\frac{\sigma_i x_i^0}{\lambda_i} e^{\lambda_i t_{budding}} - \frac{\sigma_i x_i^0}{\lambda_i} + 1 \right)^{\frac{1}{\sigma_i}}$ in the case of the protein logistic growth and $A_i = \exp \left(\left(\beta_i t_{budding} + (\beta_i / \delta_i - x_i^0) (e^{-\delta_i t_{budding}} - 1) \right) / \delta_i \right)$ in the case of the constant protein source.

By repeating the arguments in section 3.2, at least one p_{ij} has to be positive. Let us denote it

$p_{i_0 j_0}$. One can now write $p_{i_0 j_0} = c - \sum_{i \neq i_0, j \neq j_0} p_{ij}$ and formulate a new optimization problem as follows:

$$\min_{0 \leq p_{ij} \leq 1} \left(\sum_{i \neq i_0} T_i^0 A_i^{\sum_{\{j(i,j) \neq (i_0, j_0)\}} \varepsilon^j p_{ij}} + T_{i_0}^0 A_{i_0}^{\left(c - \sum_{i \neq i_0, j \neq j_0} p_{ij} \right) \varepsilon^{j_0}} \right). \quad (29)$$

As in section 3.2, this is a convex optimization problem for which the KKT point is the unique minimum. The KKT derivation can be found in Appendix F. There we get

$$p_{ij} (p_{ij} - 1) (\alpha_i \varepsilon^j - \alpha_0) = 0 \quad (30)$$

where $\alpha_0 \doteq T_{i_0}^0 A_{i_0}^{\left(c - \sum_{j \neq j_0} p_{ij} \right) \varepsilon^{j_0}}$, $\alpha_i \doteq T_i^0 A_i^{\sum_{j \neq j_0} \varepsilon^j p_{ij}} \ln A_i$. The conclusion from Eq. (30) is that for each protein i , there is one and only one epitope type j , for which $0 < p_{ij} < 1$ (i.e., the

rest are either zero or one). Arranging the epitopes according to their immunodominance, all potential epitopes with indexes greater than j will be removed. This result is supported by numerical simulations (Fig. 6). Furthermore, the numerical simulations show that as the immunodominance factor ε increases, the epitope distribution between the early and late proteins becomes more uniform (Fig. 7), as is obviously expected given the lesser importance of the total epitope number.

7 INTRA-HOST AND INTER-HOST OPTIMIZATION

We have here presented an optimization analysis. However, the optimal epitope distribution was computed assuming a given HLA haplotype. Within the human population there is a large HLA polymorphism (Gog et al., 2003). Even if a virus is optimized within a given host, upon the transition to a new host with different HLA alleles, a different epitope distribution may be required, since the peptides that can be epitopes differ among hosts. Moreover, escape mutations that confer a within host advantage, such as avoidance of clearance by CTLs, may have a between host disadvantage, such as a reduced inter-host transmission rate (Fryer et al., 2010; Grenfell et al., 2004).

Still some mutations can be transferred between hosts. There are two possibilities to explain the evolutionary transfer of such escape mutations:

- A. Mutations affecting the pre-processing stage can be transferred between hosts, since they are not affected by the HLA polymorphism. We and others have shown that such mutations are frequent (Maman et al., 2011).
- B. While there is a large polymorphism in the HLA locus, most MHC molecules have a “generic signature” within super-types and even partially between super-types. These super-types often have frequencies of over 20 % in a given population. Thus mutations affecting a given allele are often advantageous in other alleles.

If epitopes are removed by preventing the cleavage or TAP binding process, then within-host evolutionary advantage can be transferred to other hosts, unless obviously the same mutation reduces the transmission rate. If a systematic decrease in epitope number is to be maintained over evolutionary periods, it should be obtained at the epitope pre-processing level (i.e. cleavage and TAP). We thus checked if the fraction of preprocessed epitopes is indeed lower in early expression genes. We followed the division of early and late proteins described in (Vider-Shalit et al., 2009b) and computed for each sequence the fraction of epitopes cleaved and binding TAP, ignoring the MHC binding (Fig. 8). For most viruses, early expressed proteins indeed have a lower fraction of preprocessed peptides than late proteins. Moreover, in some viruses, such as Polio and some adenoviruses, the entire difference between the early and late proteins epitope densities is induced by the pre-processing stage. One can thus conclude that at least some viruses do manage to transmit a part of the within host evolutionary advantage to a population level evolutionary advantage.

8 DISCUSSION

We have here presented for the first time an optimization analysis and the corresponding bioinformatic results showing that viruses have systematically evolved over long periods to hide epitopes from early expressed proteins. Assuming that T cell clones populations evolve over a much longer period than the life cycle of a virus in an infected cell, early expressed genes should be hidden for two reasons. At the short time scale of the life cycle of a virus, the early expression of epitopes allows existing T cells to destroy infected cells before budding can occur. At the longer time scale of the T cell clone growth, the early expression of epitopes can boost T cell clones, immediately after the host infection. At the first stage of the disease both elements are important, and a long time after the infection, only the first

element should be considered. The difference between the two periods can be observed in the more uniform expected epitope distribution in the asynchronous than in the synchronous model.

Another important factor is the host cell expected survival time. If this time is long compared to the difference in the viral proteins expression rate, the effect of the viral protein expression rate on the epitope number is limited. In such cases the main element affecting the optimal epitope number is the average viral protein copy number. Finally, immunodominance should be considered. If a single epitope affects detection, then this epitope will be removed with no considerable change in the total epitope number.

We have here considered simple models. Our results can obviously be enlarged to contain more complex functions or constraints. We do not expect to see significant differences in the results for more complex functions if the general behavior of the protein copy number grows and saturates. However, the results can be affected significantly if some proteins reach a maximum and decrease back to their baseline. More complex constraints can also change the results. An important caveat is the difference between the primary and the secondary response. We have assumed uniform initial T cell clone sizes. In a secondary response, a single T cell clone may be dominant. In such a case, this may be the most significant factor affecting the optimal epitope density.

The presented results show the first systematic analysis of the differential number of epitopes on viral proteins. These results are in good agreement with our recent findings that the epitope density varies among different proteins in viruses as well as in their human host (Almani et al., 2009; Ginodi et al., 2008; Maman et al., 2011; Vider-Shalit et al., 2009b). In future research, we hope to extend this analysis to a more precise computation of the optimal epitope density incorporating a large number of factors.

Bibliography

- Aebischer T, Moskophidis D, Rohrer UH, Zinkernagel RM, Hengartner H. In vitro selection of lymphocytic choriomeningitis virus escape mutants by cytotoxic T lymphocytes. *Proc Natl Acad Sci U S A*. 1991; 88:11047–51. [PubMed: 1722316]
- Almani M, Raffaelli S, Vider-Shalit T, Tsaban L, Fishbain V, Louzoun Y. Human self-protein CD8+ T-cell epitopes are both positively and negatively selected. *Eur J Immunol*. 2009; 39:1056–65. [PubMed: 19291702]
- Anderson, RM.; May, RM. *Infectious diseases of humans: dynamics and control*. Oxford University Press; New York, New York, USA: 1991.
- Bowen DG, Walker CM. Mutational escape from CD8+ T cell immunity: HCV evolution, from chimpanzees to man. *J Exp Med*. 2005; 201:1709–14. [PubMed: 15939787]
- Braciale TJ, Braciale VL. Antigen presentation: structural themes and functional variations. *Immunol Today*. 1991; 12:124–9. [PubMed: 1711855]
- Fryer HR, Frater J, Duda A, Roberts MG, Phillips RE, McLean AR. Modelling the Evolution and Spread of HIV Immune Escape Mutants. *PLoS Pathogens*. 2010; 6:e1001196. [PubMed: 21124991]
- Ginodi I, Vider-Shalit T, Tsaban L, Louzoun Y. Precise score for the prediction of peptides cleaved by the proteasome. *Bioinformatics*. 2008; 24:477–83. [PubMed: 18216070]
- Gog JR, Rimmelzwaan GF, Osterhaus AD, Grenfell BT. Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza A. *Proc Natl Acad Sci U S A*. 2003; 100:11143–7. [PubMed: 12954978]
- Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*. 2004; 303:327. [PubMed: 14726583]
- Guldberg, CM. *Abhandlungen aus den jahren 1864, 1867, 1879*. W. Engelmann; Leipzig: 1899. Untersuchungen über die chemischen affinitäten.

- Haase AT. Population biology of HIV-1 infection: viral and CD4+ T cell demographics and dynamics in lymphatic tissues. *Annual review of immunology*. 1999; 17:625–656.
- Harty JT, Tvinnereim AR, White DW. CD8+ T cell effector mechanisms in resistance to infection. *Annu Rev Immunol*. 2000; 18:275–308. [PubMed: 10837060]
- Janeway, C. *Immunobiology: the immune system in health and disease*. Garland Pub; New York: 2001.
- Karush, W. *Minima of Functions of Several Variables with Inequalities as Side Constraints*. Dept. of Mathematics, Illinois, Univ. of Chicago; Chicago: 1939.
- Kleijmeer MJ, Kelly A, Geuze HJ, Slot JW, Townsend A, Trowsdale J. Location of MHC-encoded transporters in the endoplasmic reticulum and cis-Golgi. *Nature*. 1992; 357:342–4. [PubMed: 1589036]
- Kuhn, HW.; Tucker, AW. *Nonlinear programming*. In: Neyman, J., editor. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press; 1951. p. 481–492.
- Louzoun Y, Vider T. Score for Proteasomal Peptide Production Probability. *Immunology*. 2004; 1
- Louzoun Y, Vider T, Weigert M. T-cell epitope repertoire as predicted from human and viral genomes. *Mol Immunol*. 2006; 43:559–69. [PubMed: 15927255]
- Luenberger, DG.; Ye, Y. *Linear and nonlinear programming*. Springer; New York: 2008.
- Maman Y, Blancher A, Benichou J, Yablonka A, Efroni S, Louzoun Y. Immune-induced evolutionary selection focused on a single reading frame in overlapping hepatitis B virus proteins. *J Virol*. 2011; 85:4558–66. [PubMed: 21307195]
- McMichael AJ, Phillips RE. Escape of human immunodeficiency virus from immune control. *Annu Rev Immunol*. 1997; 15:271–96. [PubMed: 9143689]
- Metchnikoff, E.; Binnie, FG. *Immunity in infective diseases*. University Press; Cambridge: 1905.
- Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*. 1994; 152:163–75. [PubMed: 8254189]
- Peters B, Bulik S, Tampe R, Van Endert PM, Holzhutter HG. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J Immunol*. 2003; 171:1741–9. [PubMed: 12902473]
- Pircher H, Moskophidis D, Rohrer U, Burki K, Hengartner H, Zinkernagel RM. Viral escape by selection of cytotoxic T cell-resistant virus variants in vivo. *Nature*. 1990; 346:629–33. [PubMed: 1696684]
- Rehermann B, Nascimbeni M. Immunology of hepatitis B virus and hepatitis C virus infection. *Nature Reviews Immunology*. 2005; 5:215–229.
- Rimmelzwaan GF, Berkhoff EG, Nieuwkoop NJ, Smith DJ, Fouchier RA, Osterhaus AD. Full restoration of viral fitness by multiple compensatory co-mutations in the nucleoprotein of influenza A virus cytotoxic T-lymphocyte escape mutants. *J Gen Virol*. 2005; 86:1801–5. [PubMed: 15914859]
- Rock KL, York IA, Saric T, Goldberg AL. Protein degradation and the generation of MHC class I-presented peptides. *Adv Immunol*. 2002; 80:1–70. [PubMed: 12078479]
- Sebzda E, Mariathasan S, Ohteki T, Jones R, Bachmann MF, Ohashi PS. Selection of the T cell repertoire. *Annu Rev Immunol*. 1999; 17:829–74. [PubMed: 10358775]
- Soderholm J, Ahlen G, Kaul A, Frelin L, Alheim M, Barnfield C, Liljestrom P, Weiland O, Milich DR, Bartenschlager R, Sallberg M. Relation between viral fitness and immune escape within the hepatitis C virus protease. *Gut*. 2006; 55:266–274. [PubMed: 16105887]
- Steven NM, Annels NE, Kumar A, Leese AM, Kurilla MG, Rickinson AB. Immediate early and early lytic cycle proteins are frequent targets of the Epstein-Barr virus-induced cytotoxic T cell response. *The Journal of experimental medicine*. 1997; 185:1605. [PubMed: 9151898]
- Sykulev Y, Cohen RJ, Eisen HN. The law of mass action governs antigen-stimulated cytolytic activity of CD8+ cytotoxic T lymphocytes. *Proc Natl Acad Sci U S A*. 1995; 92:11990–2. [PubMed: 8618829]

- Townsend A, Bodmer H. Antigen recognition by class I-restricted T lymphocytes. *Annu Rev Immunol.* 1989; 7:601–24. [PubMed: 2469442]
- van den Broek MF, Hengartner H. The role of perforin in infections and tumour surveillance. *Exp Physiol.* 2000; 85:681–5. [PubMed: 11187963]
- van Rompay KKA, Dailey PJ, Tarara RP, Canfield DR, Aguirre NL, Cherrington JM, Lamy PD, Bischofberger N, Pedersen NC, Marthas ML. Early short-term 9-[2-(R)-(phosphonomethoxy) propyl] adenine treatment favorably alters the subsequent disease course in simian immunodeficiency virus-infected newborn rhesus macaques. *The Journal of Virology.* 1999; 73:2947.
- Verhulst PF. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance mathématique et physique.* 1838; 10:113–121.
- Vider-Shalit T, Raffaeli S, Louzoun Y. Virus-epitope vaccine design: informatic matching the HLA-I polymorphism to the virus genome. *Mol Immunol.* 2007a; 44:1253–61. [PubMed: 16930710]
- Vider-Shalit T, Fishbain V, Raffaeli S, Louzoun Y. Phase-dependent immune evasion of herpesviruses. *J Virol.* 2007b; 81:9536–45. [PubMed: 17609281]
- Vider-Shalit T, Almani M, Sarid R, Louzoun Y. The HIV hide and seek game: an immunogenomic analysis of the HIV epitope repertoire. *Aids.* 2009a; 23:1311–8. [PubMed: 19550286]
- Vider-Shalit T, Sarid R, Maman K, Tsaban L, Levi R, Louzoun Y. Viruses selectively mutate their CD8+ T-cell epitopes--a large-scale immunomic analysis. *Bioinformatics.* 2009b; 25:i39–44. [PubMed: 19478014]

Appendix A

Size of Immune Repertoire (*SIR*) score is the normalized epitope density of a protein or an organism (Ginodi et al., 2008; Louzoun and Vider, 2004; Louzoun et al., 2006; Vider-Shalit et al., 2007a; Vider-Shalit et al., 2007b; Vider-Shalit et al., 2009a; Vider-Shalit et al., 2009b). This score is calculated as follow:

1. The number of predicted CTL epitopes from a sequence was computed by applying a sliding window of nine amino acids, and computing for each nine-mer and its two flanking residues three algorithms: a proteasomal cleavage algorithm (Ginodi et al., 2008), a TAP binding algorithm developed by Peters et al. (Peters et al., 2003) and the BIMAS MHC binding (Parker et al., 1994) algorithm. The algorithms' quality was systematically validated vs. epitope databases and was found to induce low FP and FN error rates.
2. The *SIR* score was defined as the ratio between the computed CTL epitope density (fraction of nine-mers that were predicted to be epitopes) and the epitope density expected within the same number of random nine-mers.

The random nine-mers were taken from a long random peptide built, using the amino acid distribution calculated over the sequences of all fully sequences viruses available, and taking into account the correlation between the frequencies of neighboring amino acids in these virus (Vider-Shalit et al., 2007b). For example, assume a hypothetical sequence of 1,008 amino acids (1,000 nine-mers) containing 15 HLA A*0201 predicted epitopes. If the average epitope density of HLA A*0201 in a large number of random proteins with an amino acid distribution typical to viruses was 0.01 (i.e. 10 epitopes in 1,000 nine-mers), then the *SIR* score of the sequence for HLA A*0201 would be 1.5 (15/10).

3. Finally, the average *SIR* score of a protein was defined as the average of the *SIR* scores for each HLA allele studied, weighted by the allele's frequency in the average human population.

The computation of the *SIR* scores can be performed through our web-server at <http://peptibase.cs.biu.ac.il/index.html>.

Appendix B

We here show that in the case of the asynchronous phase none of the constraints are active.

Lemma 1

The unique possible solution of the minimization problem $\mu(\mathbf{p}^*) = \min \sum_i \mathbf{A}_i \cdot (\mathbf{p}_i)^2$ is $(p_1^*, p_2^*, \dots, p_n^*)$ where $\forall_{1 \leq i \leq n} \cdot p_i^* > 0$.

Proof

The minimization of $\mu(\mathbf{p}^*) = \min \sum_i \mathbf{A}_i \cdot (\mathbf{p}_i)^2$ is equivalent to the minimization of

$$\mu(\bar{\mathbf{p}}) = \frac{1}{2} \bar{\mathbf{p}}^T (2\mathbf{A} + 2\mathbf{A}_{i_0} \bar{\mathbf{b}} \bar{\mathbf{b}}^T) \bar{\mathbf{p}} - 2\mathbf{A}_{i_0} \bar{\mathbf{c}} \bar{\mathbf{b}}^T \bar{\mathbf{p}} + \mathbf{c} \mathbf{A}_{i_0}$$

where

$$\mathbf{A} = \begin{pmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_n \end{pmatrix}, \quad \bar{\mathbf{b}} = \begin{pmatrix} 1 \\ \dots \\ \dots \\ 1 \end{pmatrix}$$

By definition, $2(\mathbf{A} + \mathbf{A}_{i_0} \bar{\mathbf{b}} \bar{\mathbf{b}}^T)$ is the Hessian

$$2(\mathbf{A} + \mathbf{A}_{i_0} \bar{\mathbf{b}} \bar{\mathbf{b}}^T) = 2 \begin{pmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_n \end{pmatrix} + 2\mathbf{A}_{i_0} \begin{pmatrix} 1 & \dots & \dots & 1 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 1 & \dots & \dots & 1 \end{pmatrix}$$

The first and second matrices are positive definite, thus the Hessian is positive definite, leading to a necessary and sufficient condition for the presence of a minimum point.

Appendix C

Let us rewrite problem(15):

$$L(\mathbf{p}, \eta, \nu) = \sum_i T_i^0 A_i^{p_i} - \nu \left(\sum_i p_i - c \right) - \sum_i \eta_i p_i \quad (31)$$

The KKT conditions for problem (31) are:

$$\frac{\partial L}{\partial p_i} = T_i^0 A_i^{p_i} \ln A_i - \nu - \eta_i = 0, \quad 1 \leq i \leq n, \quad (31a)$$

$$\frac{\partial L}{\partial \eta_i} = -p_i \leq 0, \quad 1 \leq i \leq n, \quad (31b)$$

$$\frac{\partial L}{\partial v} = -\sum_i p_i + c \leq 0, \quad (31c)$$

$$\eta_i p_i = 0, \quad 1 \leq i \leq n, \quad (31d)$$

$$v \left(\sum_i p_i - c \right) = 0. \quad (31e)$$

$$\eta_i \geq 0, \quad 1 \leq i \leq n; v \geq 0 \quad (31f)$$

Following the constraint $\sum_i p_i \geq c$, at least one of the p_i has to be positive (and the appropriate η_i will obviously be 0). Let us denote this gene as i_0 , $p_{i_0} \neq 0$. One can use i_0 to extract the value of v to be: $v = T_{i_0}^0 A_{i_0}^{p_{i_0}} \ln A_{i_0}$. This value is always positive, so that $v > 0$ is an active constraint. In other words, as one could expect, if the total amount of epitopes is limited to be higher than c , it will be precisely c . As in the asynchronous case, we can write $p_{i_0} = c - \sum_{i \neq i_0} p_i$ and formulate a new optimization problem:

$$\min_{p_i \geq 0} \left(\sum_{i \neq i_0} T_i^0 A_i^{p_i} + T_{i_0}^0 A_{i_0}^{c - \sum_{i \neq i_0} p_i} \right) \quad (32)$$

with the following solution for the active constraints:

$$\eta_i = T_i^0 A_i^{p_i} \ln A_i - T_{i_0}^0 A_{i_0}^{c - \sum_{i \neq i_0} p_i} \ln A_{i_0}, \quad 1 \leq i \leq n \wedge i \neq i_0 \quad (33)$$

Appendix D

We here show that the optimization results for the synchronous model are similar whether all constraints (31b) are active or not. If some constraints are active, we denote in the language of Eq. (16): $\eta_i \neq 0$, $1 \leq i \leq n \wedge i \neq i_0$.

There are $2^n - 2$ possible ensembles of active constraints. Once the set I_a of active constraints contributing to the zero solution $p_i = 0$ for all $i \notin I_a$ is found, the solution of (14) reduces to the solution of (32) with a lower dimension. The existence of $p_i = 0$ solutions raises the possibility that viruses hide the epitopes of some proteins and maintain their survival fit through the optimization of other proteins. The general solution of (32) is:

$$p_i = \begin{cases} 0 & ; i \in I_a \\ \frac{1}{\ln A_i} \left[\ln Z_i + \left(c - \sum_{j \in I_{na}} \frac{\ln Z_j}{\ln A_j} \right) / \sum_{j \in I_{na}} \frac{1}{\ln A_j} \right] & ; i \in I_{na} \end{cases} \quad (34)$$

According to (34), the condition for positive p_i (i.e. the I_{na}) can be formulated as $\forall i \notin I_a$,

$\ln Z_i + \left(c - \sum_{j \in I_{na}} \frac{\ln Z_j}{\ln A_j} \right) / \sum_{j \in I_{na}} \frac{1}{\ln A_j} > 0$ (note that $\forall i; A_i > 1$), leading to a condition for the presence of protein i in I_{na} :

$$\sum_{j \in I_{na}} \frac{\ln Z_i - \ln Z_j}{\ln A_j} + c > 0 \Rightarrow \sum_{j \in I_{na}} \frac{q_j - q_i}{\ln A_j} + c > 0, \quad (35)$$

where $q_j = \ln(T_j^0 \ln(A_j))$, based on the definition of Z_i .

In order to compute I_{na} , based on condition (35), let us note that following the definition of A_j : $\forall j: \ln A_j > 0$. Thus, only the differences $q_j - q_i$ affect the sign of the summed elements. The factor q_i can be interpreted as a measure of the contribution of the gene i to the total

cost. In the first step we include in the set I_{na} all the indexes i , fulfilling: $\sum_{j=1}^n \frac{q_j - q_i}{\ln A_j} + c > 0$ (note that the sum here is on all proteins), excluding from the sum in each step the indexes that do not obey this condition, and assigning them to the set of active constraints I_a . We here prove that the protein set defined by the first step of the algorithm is a subset of I_a .

Lemma 2

The set I_a of all the indexes obeying $\sum_{j=1}^n \frac{q_j - q_i}{\ln A_j} + c \leq 0$ is a subset of I_a .

Proof

We first sort $q_1 \dots q_n$ in increasing order. Let us denote the sorted list as q^s and its indexes as

j^s . There exists a first index in j^s denoted by $j_{\min}^s$, for which. $\sum_{j=1}^n \frac{q_j - q_{j_{\min}^s}}{\ln A_j} + c \leq 0$. Otherwise,

for all i : $\sum_{j=1}^n \frac{q_j - q_i}{\ln A_j} + c > 0$, and $I_a = \emptyset \subseteq I_a$. Since the list q^s is sorted, for all

$$k > j_{\min}^s: \sum_{j=1}^n \frac{q_j - q_k}{\ln A_j} + c = \sum_{j=1}^n \frac{q_j - q_{j_{\min}^s}}{\ln A_j} + \sum_{j=1}^n \frac{q_{j_{\min}^s} - q_k}{\ln A_j} + c \leq 0.$$

Thus $\tilde{I}_a = \{k | k \geq j_{\min}^s\}$.

Let $k1$ be the member of I_a with the minimal value of q_k , for all k with higher values of q_k ,

$$\sum_{j \in I_{na}} \frac{q_j - q_k}{\ln A_j} + c \leq \sum_{j \in I_{na}} \frac{q_j - q_{k1}}{\ln A_j} + c \leq 0 \Rightarrow k \in I_a.$$

Assume that $k1 > j_{\min}^s$, then

$$0 \geq \sum_{j=1}^n \frac{q_j - q_{j_{\min}^s}}{\ln A_j} + c = \left[\sum_{j=1}^{k1-1} \frac{q_j - q_{j_{\min}^s}}{\ln A_j} + c \right] + \sum_{j=k1}^n \frac{q_j - q_{j_{\min}^s}}{\ln A_j} = \left[\sum_{j \in I_{na}} \frac{q_j - q_{j_{\min}^s}}{\ln A_j} + c \right] + \sum_{j=k1}^n \frac{q_j - q_{j_{\min}^s}}{\ln A_j} > 0.$$

The first part of the inequality represents the fact that $k1 > j_{\min}^s$ and thus $j_{\min}^s \in I_{na}$ and the second part results from the fact that the q_k values are sorted.

In the second step, we check whether the constructed set I_a satisfies the condition (35). If it does not, we update it returning to the first step until the condition is satisfied. This algorithm provides the optimal solution, and its cost is $O(n^2)$. The flow chart of the algorithm is presented in Figure. D.1.

Appendix E

Another possible model is that the bulk of virus release occurs while a cell is still alive and CTL induced lysis occurs significantly after the onset of budding. In such a case the mean value of virions produced after the budding and before host cell lysis has to be optimized, and not the survival probability before budding.

We assume that the T cell clones population has reached a steady state and use the constant growth model to describe protein dynamics:

$$\begin{cases} \dot{x}_i = \beta_i - \delta_i x_i \\ T_i = T_i^0 p_i \int_0^{t_{death}} x_i dt, \quad 1 \leq i \leq n \end{cases} \quad (36)$$

where t_{death} is a time of cell lysis.

The solution of these dynamics is

$$\begin{cases} x_i = \frac{\beta_i}{\delta_i} + \left(x_0^i - \frac{\beta_i}{\delta_i} \right) e^{-\delta_i t} \\ T_i = \frac{T_i^0 p_i}{\delta_i} \left[\beta_i t_{death} + \left(\frac{\beta_i}{\delta_i} - x_0^i \right) e^{-\delta_i t_{death}} - \left(\frac{\beta_i}{\delta_i} - x_0^i \right) \right], \quad 1 \leq i \leq n \end{cases} \quad (37)$$

The survival probability until time t can be defined as before:

$$\mu(t) = \int_{t_{budding}}^t dt' \sum_i p_i x_i(t') T_i = \sum_i p_i T_i \cdot \left[\frac{\beta_i}{\delta_i} t + \frac{A_i}{\delta_i} e^{-\delta_i t} - \frac{A_i}{\delta_i} \right]; A_i = \frac{\beta_i}{\delta_i} - x_0^i. \quad (38)$$

However, the optimization problem is very different. The function optimized now is

$$\max_{\sum_i p_i \geq c, p \geq 0} E(\text{virion number}) \propto \max E(t_{death} - t_{budding}) = \max_{\sum_i p_i \geq c, p \geq 0} \int_{t_{budding}}^{\infty} t e^{-\mu(t)} dt. \quad (39)$$

Let us set $\delta_1 \geq \delta_2 \geq \dots \geq \delta_n$ and use the estimation: $e^{-\delta_i t} = \begin{cases} 1 - \delta_i t, & t \leq 1/\delta_i \\ 0, & t > 1/\delta_i \end{cases}$ and the

notation $s = \sum_i p_i T_i \frac{\beta_i}{\delta_i}$, $d_i = p_i T_i \frac{A_i}{\delta_i}$, to get

$$E(t_{death} - t_{budding}) = \int_{t_{budding}}^{\infty} t e^{-st + \sum_i d_i (1 - e^{-\delta_i t})} dt \approx \int_{t_{budding}}^{\infty} t e^{-st + \sum_i d_i \cdot \min\{\delta_i t, 1\}} dt. \quad (40)$$

We denote the time to saturation of the first protein saturating after budding by t_1 : $t_1 = 1/\delta_{\arg\min\{i \geq t_{budding}\}}$, and by t_i , $2 \leq i \leq m$ expression times that occur after t_1 . We rename the set

$\{i\}$ to be the new set of indexes (that is the new $i=1$ actually is equal to $\arg\min\{i \geq t_{budding}\}$ in the old set of i). Note that all proteins are expressed before the budding time. The only difference between proteins is the time to saturation. Setting

$s_i = -s + \sum_{j=i+1 \dots m} d_j \delta_j$, $k_i = \exp\left(\sum_{j=0 \dots i} d_j\right)$, we get the estimation for the expected number of budding virions:

$$\begin{aligned} E(t_{death} - t_{budding}) &= \frac{1}{s_0} \left(e^{s_0 t_0} \left(t_0 - \frac{1}{s_0} \right) - e^{s_0 t_{budding}} \left(t_{budding} - \frac{1}{s_0} \right) \right) \\ &+ \sum_{i=1}^{m-1} e^{k_i} \left(e^{s_{i+1} t_{i+1}} \left(t_{i+1} - \frac{1}{s_{i+1}} \right) - e^{s_i t_i} \left(t_i - \frac{1}{s_i} \right) \right) \\ &+ \frac{1}{s} \left[\exp(-s t_m) \cdot \left(t_m + \frac{1}{s} \right) \right] \end{aligned} \quad (41)$$

This expectation can be maximized numerically to provide results that are much more sensitive to the total protein copy number than the expression timing (results not shown). Such a result is expected since in this model budding is a long term process and the precise details of the initial stages of the process are not important. Thus in such viruses, we expect no correlation between the protein expression time and the number of epitopes.

Appendix F

The Lagrangian of Eq. (29) is

$$L(\mathbf{p}, \eta, \theta) = \sum_{i \neq i_0} T_i^0 A_i \left(\sum_{j \in \{i, j\} \neq \{i_0, j_0\}} \varepsilon^j p_{ij} \right) + T_{i_0}^0 A_{i_0} \left(c - \sum_{i \neq i_0, j \neq j_0} p_{ij} \right) \varepsilon^{j_0} - \sum_{i,j} \eta_{ij} p_{ij} + \sum_{i,j} \theta_{ij} (p_{ij} - 1). \quad (42)$$

In this case, the KKT conditions are:

$$\frac{\partial L}{\partial p_{ij}} = T_i^0 A_i \sum_{j \neq j_0} \varepsilon^j p_{ij} \varepsilon^j \ln A_i - T_{i_0}^0 A_{i_0} \left(c - \sum_{j \neq j_0} p_{ij} \right) \varepsilon^{j_0} \varepsilon^j \ln A_{i_0} - \eta_{ij} + \theta_{ij} = 0, \quad (42a)$$

$$\frac{\partial L}{\partial \eta_{ij}} = -p_{ij} \leq 0, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m, \quad (42b)$$

$$\frac{\partial L}{\partial \theta_{ij}} = p_{ij} - 1 \leq 0, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m, \quad (42c)$$

$$\eta_{ij} p_{ij} = 0, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m, \quad (42d)$$

$$\theta_{ij} (p_{ij} - 1) = 0, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m, \quad (42e)$$

$$\eta_{ij} \geq 0, \quad \theta_{ij} \geq 0, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m. \quad (43f)$$

Denoting $\alpha_0 \doteq T_{i_0}^0 A_{i_0} \left(c - \sum_{j \neq j_0} p_{ij} \right) \varepsilon^{j_0} \ln A_{i_0}$, $\alpha_i \doteq T_i^0 A_i^{\sum_{j \neq j_0} \varepsilon^j p_{ij}}$, multiplying the condition (42a) by $p_{ij} (p_{ij} - 1)$ and applying (42d), (42e), the condition (42a) takes the form:

$$\frac{\partial L}{\partial p_{ij}} = p_{ij} (p_{ij} - 1) (\alpha_i \varepsilon^j - \alpha_0) = 0 \quad (43)$$

Figure 1. Observed *SIR* score of early vs. late proteins

The data is shown for 24 viruses for multiple HLA alleles (31 alleles). Each column represents the ratio between the difference of the late and early *SIR* score to their

sum $\left(\frac{SIR(late) - SIR(early)}{SIR(late) + SIR(early)} \right)$. For most HLA\virus combinations, the ratio is more than zero, indicating a larger number of epitopes in late expressed proteins.

Figure 2. Schematic drawing of viral protein and T cell dynamics

The upper right drawing represents the viral protein copy number for two proteins. The full line is an early protein, while the dashed line is a late protein. The early protein induces a higher death rate probability per epitope than the late one (upper left drawing). Thus in order to maximize the survival probability, the number of epitopes in early proteins should be minimized. The central drawing represents the T cell dynamics of clones reacting either to the early or to the late proteins. These dynamics occur over a scale of approximately 20 days, and have three phases: the early growth phase, the decay phase and a possible latent phase. In each such phase many viral life cycles occur. Most of them are not synchronized one with each other. Since we do not know which period of the disease is most critical for the virus, we solve the optimization for the first and last stages independently.

Figure 3. Optimal number of epitopes as a function of the expression rate in asynchronous model. (A) Logistic growth

The virus budding time is normalized to 1. The early stage optimal solution is presented for a set of 10 proteins, each with a different expression rate: $\lambda_i = 3 - 0.3(i - 1)$; $1 \leq i \leq 10$. Two saturation levels are considered: a constant saturation level (full line): $\lambda_i/\sigma_i = 1$ and a saturation level that is inversely proportional to expression rate (dashed line): $(\lambda_i/\sigma_i = 1/\lambda_i)$. The total epitope number is $C = 5$. The initial lymphocyte clone size was arbitrarily set to $T_i^0 = 1$. The initial proteins concentration was $x_i^0 = 0.1$. The optimization was solved analytically according to expression (12). In both cases, early expressed proteins have a lower expected number of epitopes. **(B) Constant growth model.** The same analysis was performed for the same values of budding time, initial concentrations and total epitope number. The expression rates in this case are defined by $\beta_i = 3 - 0.3(i - 1)$; $1 \leq i \leq 10$ and the saturation levels which in this case are defined by β_i/δ_i are: $\beta_i/\delta_i = 1$ for a constant saturation level (full line), $\beta_i/\delta_i = 1/\beta_i$ for a saturation level that is inversely proportional to the expression rate (dashed line). The results are similar to the logistic growth with slightly more weight given to the saturation level.

Figure 4. Optimal number of epitopes as a function of the expression rate in synchronous model
 The presented results follow the assumptions used in Figure 3 for the asynchronous model. As in Figure 3, the left drawing (A) represents logistic growth and the right drawing (B) represents constant growth. The optimization was solved using the algorithm described in Figure D.1 (see Appendix D), as well as with the Matlab “fmincon” function with similar results. The results follow the trend of the asynchronous stage, with the exception that the epitope number can be zero for some of proteins.

Figure 5. Effect of T cells' multiple generations

The presented results follow the assumptions used in Figure 3, drawing (B) for the asynchronous model. The drawings (A) and (B) represent the growth with a constant saturation level ($\beta_i/\delta_i = 1$) and the growth for a saturation level that is inversely proportional to expression rate ($\beta_i/\delta_i = 1/\beta_i$) respectively. In each drawing the grey line stands for the optimal solution for T cells response in the first generation (see Figure 3, drawing (B)). The dashed black line stands for the optimal solution for T cell response in second generation and the grey dotted line – for the third T cells generation. In the two cases, one can see that the trend of distribution is maintained, but the distribution becomes more flat.

Figure 6. Epitope presentation probability versus immunodominance

The figure represents the optimal distribution of different epitopes for each protein (each protein is a different subplot), which are sorted inversely to their expression time (from late to early), taking into account the immunodominance. The immunodominance factor is $\varepsilon=1.2$. The results are computed for the same parameters as in Figure 3B. In each subplot (protein), the epitopes are sorted according to their affinity in increasing order, and the probability of each epitope to be presented is shown. The results show that the virus only hides the high affinity epitopes.

Figure 7. Optimal number of epitopes as a function of the expression rate in the model with immunodominance

The results are presented with the assumptions in Figure 3, drawing (B) for different values of the immunodominance factor $\varepsilon=1, 1.2, 1.4, 1.6, 1.8$. The drawings (A) and (B) represent the growth for a constant saturation level ($\beta_i/\delta_i=1$) and for a saturation level that is inversely proportional to expression rate ($\beta_i/\delta_i=1/\beta_i$) respectively. In the two cases, the epitope number distribution becomes more uniform with increasing immunodominance factor.

Figure 8. The difference between the fraction of preprocessed epitopes in early and late expressed genes

For most viruses, early expression proteins have less preprocessed peptides. In some viruses, such as Polyo and some adenoviruses, the entire difference between the epitope density in early and late proteins is induced by the preprocessing stage.

Figure D.1.
Algorithm flow chart.