Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Infection, Genetics and Evolution 20 (2013) 71–77

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Viral proteome size and CD8+ T cell epitope density are correlated: The effect of complexity on selection

Alexandra Agranovich [a,1], Yaakov Maman [b,1], Yoram Louzoun [a,*]

[a] Department of Mathematics and Gonda Brain Research Center, Bar-Ilan University, Ramat Gan 52900, Israel
[b] The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel

ARTICLE INFO

Article history:
Received 2 May 2013
Received in revised form 29 July 2013
Accepted 31 July 2013
Available online 15 August 2013

Keywords:
Evolution
Viral infections
CD8+ T cell epitopes
Mutations
Complexity

ABSTRACT

The relation between the complexity of organisms and proteins and their evolution rates has been discussed in the context of multiple generic models. The main robust claim from most such models is the negative relation between complexity and the accumulation rate of mutations.

Viruses accumulate escape mutations in their epitopes to avoid detection and destruction of their host cell by CD8+ T cells. The extreme regime of immune escape, namely, strong selection and high mutation rate, provide an opportunity to extend and validate the existing models of relation between complexity and evolution rate as proposed by Fisher and Kimura.

Using epitope prediction algorithms to compute the epitopes presented on the most frequent human HLA alleles in over 100 fully sequenced human viruses, and over 900 non-human viruses, we here study the correlation between viruses/proteins complexity (as measured by the number of proteins in the virus and the length of each protein, respectively) and the rate of accumulation of escape mutation. The latter is evaluated by measuring the normalized epitope density of viral proteins.

If the virus/protein complexity prevents the accumulation of escape mutations, the epitope density is expected to be positively correlated with both the number of proteins in the virus and the length of proteins. We show that such correlations are indeed observed for most human viruses. For non-human viruses the correlations were much less significant, indicating that the correlation is indeed induced by human HLA molecules.

1. Introduction

Evolution is driven by a combination of mutations and selection. The balance between the two is a function of the cost of mutations and the strength of selection. Mutations can have, on the one hand, an environmental advantage and, on the other hand, a fitness cost (Soderholm et al., 2006). A typical example of this dual effect is escape mutations from CD8+ T cells (CTLs) in viruses. While these mutations can lead to a higher survival probability, they often lead to a lower probability of producing functional virions. We here use bioinformatics tools to analyze the relation between the frequency of escape mutations and the organisms' complexity.

CTLs recognize virally infected cells through small (typically 8–10 amino acid long) peptides, denoted epitopes. These epitopes are presented in the binding groove of MHC class I molecules located on the surface of these cells (Williams et al., 2002). When an appropriate CTL encounters a host cell expressing such epitopes, the host cell is rapidly destroyed along with its hosted virus (Aebischer et al., 1991; Bowen and Walker, 2005; McMichael and Phillips, 1997). This leads to an evolutionary pressure on viruses to avoid this detection in order to survive and infect new cells. Peptide binding to MHC-I groove requires well-defined binding motifs. Only a few percent of the possible peptides have such a motif, limiting the number of possible epitopes in every protein to approximately 1–2% of all possible nine-mers for a given HLA allele (Yewdell, 2006). The HLA polymorphism challenges viruses with a changing environment that may result in back-and-forth (toggling) escape mutations (Delport et al., 2008) and thus limit the fixation of mutation. However, mutations in the cleavage sites of the highly conserved proteasome or in positions that binds to conserved HLA motifs can lead to a removal of epitopes at the population level. Thus, in principle, a very limited number of properly positioned mutations can completely hide a viral protein from CTLs.

Many viruses indeed acquire escape mutations in epitopes presented to CTLs (Bowen and Walker, 2005; McMichael and Phillips, 1997; Agranovich et al., 2011; Alcami, 2003; Poppema et al., 1998; Timm et al., 2004; Yates et al., 2007). These mutations have the obvious advantage of reducing the probability that a CTL would kill the virus. The balance between the fitness cost and the advantage

* Corresponding author. Tel.: +972 3 5317610; fax: +972 3 7384057.
E-mail address: louzouy@math.biu.ac.il (Y. Louzoun).
[1] These authors contributed equally to this work.

1567-1348/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.07.030

obtained by escape mutation leads to a non-uniform epitope density distribution among different viral proteins. We have recently shown, for example, that proteins expressed early in the viral life cycle have a lower epitope density than proteins expressed late in the viral life cycle (Agranovich et al., 2011; Vider-Shalit et al., 2009a, 2007), and that proteins with a low copy number have more epitopes than proteins with a high copy number (Maman et al., 2011a). Here, we study the relation between the accumulation of escape mutations and the viral complexity (as shall be further defined). Specifically, we test whether the (dis)advantage of a given mutation is determined by the mutation itself or whether it is related to the complexity of the entire protein or perhaps even the entire organism.

The relation between complexity and selection was initially studied by the pioneering work of Fisher in 1930 (Fisher, 1930). Fisher proposed that as the dimensionality of the phenotype increases, the probability of a mutation being beneficial decreases due to its pleiotropic effects and different dimensions of the phenotype. The phenotype dimensionality is defined by the number of organism's parts (phenotypic characters –denoted hereafter as n). Kimura and Orr then expanded Fisher's work and showed that Fisher underestimated the cost of complexity by not incorporating the lower fixation probability of mutations with a limited phenotypic effect following the effect of stochastic drift (Kimura, 1983; Orr, 2000). Orr (2000) further showed that the average fitness increase rate is inversely proportional to n for small and medium n and much faster for large n. In other words, the adaptation rate of complex organisms is lower than the one of simpler ones. Welch and Waxman (2003) examined the robustness of Orr's model by introducing different mechanisms (such as varying magnitude of mutations, modularity (Wagner, 1996; Wagner and Altenberg, 1996; Baatz and Wagner, 1997) and a constant mean mutational chance per phenotypic character). They showed that the relation between the complexity and adaptation rate is robust to most variations of the model. Gillespie (1994, 1984, 1983) extended Fisher's model and proposed the mutational landscape model. Orr (2003, 2002) further extended his work and found different patterns that characterize the adaptation of DNA sequences. His model was tested using single stranded DNA viruses (Rokyta et al., 2005). These studies were done in the regime of weak selection and low mutation rate.

In contrast with the above mentioned models, viral escape from immune recognition is characterized by strong selection and a high mutation rate. Most viruses budding from a given cell are destroyed and most infected cells can be cleared extremely fast in the presence of an immune response, since CTLs can induce apoptosis in infected cells within minutes (Regoes et al., 2007; Macken and Perelson, 1984) (for reviews see Yates et al., 2007; Regoes et al., 2007). The mutation rate of viruses can reach 1.e−4 mutations per base pair per replication (Sanjuan et al., 2010). We here show, using bioinformatics measurements, a direct relation between organisms' complexity (as defined by their proteins' length and their number) and their epitope density. The ratio between the epitope density and the expected epitope density based on the amino acid composition of each protein are used to estimate the accumulation of escape mutations. Viruses with a low number of proteins accumulate more escape mutations per protein than large ones, and short proteins accumulate more escape mutations than long ones, even in steady state.

Protein length and number are obviously a simplistic proxy for complexity and "phenotype dimensionality". However, in viruses where the number of proteins is highly limited, such an approximation is probably reasonable. Note that large viruses may have developed other alternatives, such as specific proteins that down-modulate MHC or MHC loading. In other words, the cost of removing mutations may be too high for them, leading them to alternative pathways to modulate the immune response.

2. Results

2.1. Evolution rate assessment using theSIR Score

The Size of Immune Repertoire (*SIR*) score for a given HLA allele (an allele that encodes for a human MHC class I molecules) is an estimate of the average normalized CTL epitope density of a given protein in this allele. Specifically, the *SIR* score of an amino acid sequence for a given HLA allele is the ratio between the predicted CTL epitope density in this sequence and the epitope density expected in a random sequence. (See Methods for detailed description). It is based on multiple Bioinformatic algorithms used to compute all stages of epitope processing and presentation, and was tested to be precise in multiple previous studies (Vider-Shalit et al., 2009a,b, 2007; Maman et al., 2011a,b; Vider-Shalit and Louzoun, 2010; Kovjazin et al., 2011).

In order to unify the score over all alleles, the *SIR* score of a protein sequence in a population is defined as the weighted average *SIR* score for all HLAs, weighted by the HLA allele frequency in that population. An average *SIR* score of less than 1 represents a sequence with less epitopes than expected; conversely, an average *SIR* score of more than 1 represents a sequence with more epitopes than expected. A schematic description of the *SIR* score is given in Fig. 1.

The SIR score of a virus is then defined as the average *SIR* score of all its proteins. Note that large and small proteins have equal weights in this analysis.

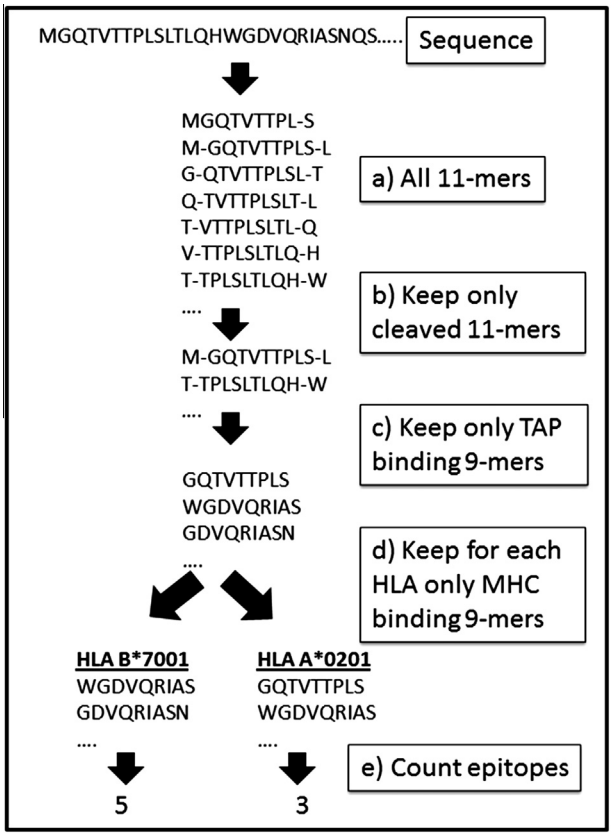

Fig. 1. Algorithm for the *SIR* score computation. Each viral protein is divided into all nine-mers and the appropriate flanking regions (a). For each nine-mer a cleavage score is computed (b). We compute a TAP binding for all nine-mers with a positive cleavage score and choose only supra-threshold peptides (c). Using the MLVO algorithm, the MHC binding scores of all TAP binding and cleaved nine-mers are computed (d). Nine-mers passing all these stages are defined as epitopes. We then compute the number of epitopes per protein per HLA allele (e).

2.2. Correlation between the number of proteins and the epitope density

The complexity of an organism can be measured by its protein number. Thus, if for complex organisms the fixation probability of escape mutation is lower, we expect a positive correlation between the average *SIR* score of an entire virus and the number of proteins in the virus. In other words, we expect viruses with fewer proteins to have a lower average epitopes density (ratio between epitope numbers and Protein lengths).

We have analyzed the proteins of all viruses (human and non-human) with at least 4 proteins with a RefSeq identifier (Pruitt et al., 2007). A full list of the studied viruses and proteins is given in the Supplementary Material.

A positive correlation was observed between the average *SIR* score of all proteins in a virus and the number of proteins in the virus, in viruses infecting humans (Spearman $R = 0.23$, $p < 0.0016$) (Fig. 2).While in general most of the viruses have average *SIR* scores lower than 1 (T test p value 9.2e−6), viruses with a small number of proteins (less than 40) are more biased toward a lower scores (average 0.87) than viruses with a large number of proteins (average 0.97).

RNA viruses tend to undergo more mutations and probably generate more escape mutations. Also, their genome size is relatively smaller as compared to DNA viruses. A simple hypothesis could have been that the protein number effect is induced by the difference between RNA and DNA viruses. We have thus performed a regression analysis on the virus type (DNA vs. RNA) and the protein number (data not shown). We found that indeed RNA viruses had a significantly lower *SIR* score than DNA viruses ($p < 0.001$). However, even the virus type was incorporated, the regression coefficient of the *SIR* score on the number of proteins was positive and significant ($p < 0.001$).

In order to test the assumption that the observed correlation is indeed a result of selection against epitope presentation, we have performed a similar analysis in viruses infecting non-human hosts. Such viruses, which have never met human HLA alleles, are not expected to accumulate escape mutations with respect to these human HLA alleles. Indeed, the Spearman correlation of the protein number and the *SIR* score in viruses infecting non human hosts is practically null (Fig. S1, $R = 0.08$, $p = 0.44$). Note that since there is some similarity between human and non-human MHC molecules, peptides that are epitopes for non-human MHC molecules are sometimes also epitopes for human MHC molecules. Moreover, since the proteasome is highly conserved among species, human and non-human hosts share a similar pool of cleaved peptides that can serve as ligands for MHC-I binding.

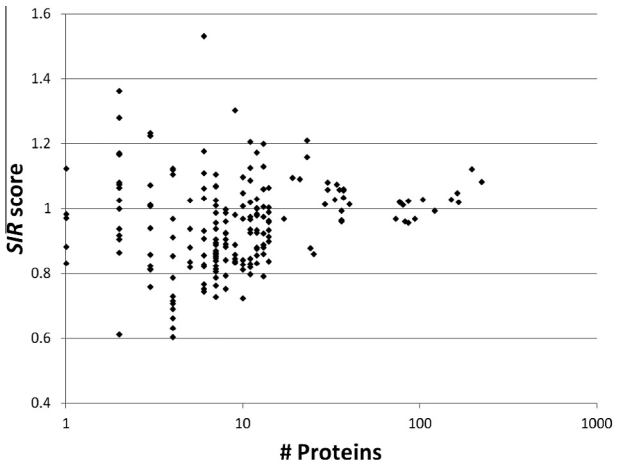

Fig. 2. Number of proteins in a virus versus its average *SIR* score. The *x* axis is the protein number for each virus. The *y* axis is the average *SIR* score. Each dot is a virus infecting a human host. Viruses with a low number of proteins have on average a low *SIR* score, while viruses with a large number of proteins have an *SIR* score of slightly less than 1. Note that some large viruses show a limited extent of selection. Furthermore, some small viruses have an SIR score higher than 1, as is expected from the low number of proteins in such viruses and the random variability in the epitope density. (Pearson, $R = 0.17$, $p < 0.015$; Spearman $R = 0.23$, $p < 0.0016$). Epitope prediction was done by the MLVO algorithm.

Many of the studied viruses are quite similar (e.g. HIV I and HIV II). The significant correlation may be the result of the similarity between viruses (adding more degrees of freedom than there actually are). In order to exclude this possibility, we repeated the same analysis grouping all viruses from the same family (e.g., all HPVs, all Herpesviruses, all Influenza viruses). The result is an even clearer correlation between the *SIR* score and the Log protein number (Spearman $R = 0.36$, $p < 0.001$) (Supplementary Material, Fig. SS4). Again, performing a similar analysis on non-human viruses yields no significant correlation ($p = 0.44$).

2.3. Correlation between protein length and epitope density

Another measurement that could reflect complexity is the protein length. While the number of proteins represents the complexity of different viruses, the protein length is correlated to the complexity within the virus. Obviously, a viral protein may include several domains and have several functions, and protein length is not completely equivalent to the complexity. However, as a first approximation, we can expect a correlation between protein length and complexity within a virus. We therefore tested the correlation between protein length and the epitope density in a virus specific manner.

If indeed selection is more active in shorter proteins than in longer proteins, we expect again a positive correlation between the proteins' length and their epitope density. If this correlation is indeed due to selection, it should only be observed more in viruses infecting human hosts, than in viruses not infecting human hosts.

We tested the correlation between *SIR* score and protein length for each virus and plotted the distribution of the correlation coefficient for human viruses (Fig. 3)). In viruses infecting human hosts, this correlation is positive for most cases (about 80% of cases), and the average Spearman correlation is 0.25 and is significantly higher than 0 (one sample T test with the average correlation per virus, $p = 3.e{-}15$).

A similar test on non-human virus produced a smaller, yet significant deviation from zero (average $R = 0.1$, $p = 0.025$, Fig. S2). When comparing the average correlation in human and non-human viruses, non-human viruses have a much lower average correlation (two sample T test with the average correlation per virus T test, $p < 0.01$). Again, the presence of some correlation between protein length and *SIR* score in non-human viruses is probably the result of the partial similarity between the MHC binding motifs among mammals. Note that for some viruses random fluctuations or other elements affecting the epitope density could induce negative correlations. Therefore, although not the ultimate factor, the organism "size" is a major factor affecting the selection against epitope presentation.

In order to ensure that the results are not due to a single family of viruses, we repeated the analysis for each family separately. This resulted in similar results (Fig. S3). One can see that for most families, the correlation coefficient is positive.

As we have shown previously, the selection acts mainly on viruses with a limited number of proteins (Fig. 2). Therefore, if indeed the correlation between the protein length and the *SIR* score is induced by selection, we expect the *SIR*-length correlation coefficients to be correlated with the number of proteins in the virus.

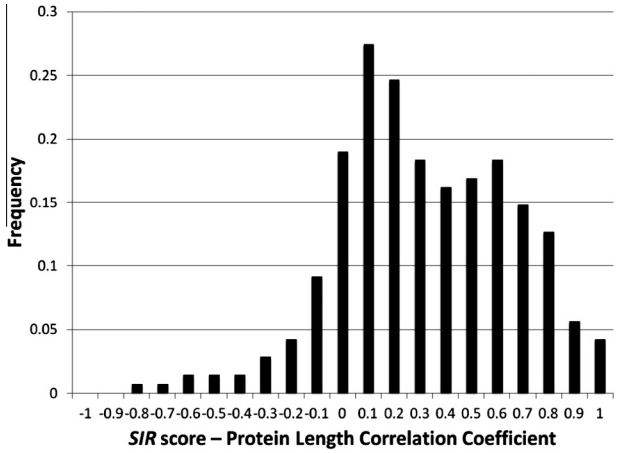

Fig. 3. Histogram of Spearman correlation coefficients for viruses infecting human hosts. We computed the correlation between the *SIR* Score and the protein length for all proteins within each human virus. The *R* values of the correlations are clearly biased toward positive values (one sample T test with the average correlation per virus, $p = 3.e{-}15$). There are practically no viruses with very negative correlations, but there are some viruses with very high positive correlations between the protein length and the *SIR* score of the protein. Epitope prediction was done by the MLVO algorithm.

Indeed, viruses with a high number of proteins have no correlation between the protein length and the *SIR* score, while practically all small viruses have such a correlation. Specifically, the correlation between the number of proteins in the virus and the *SIR*-length correlation coefficients is significant in human viruses ($R = -0.27$, $p < 1.e{-}5$) (Fig. 4). No such correlation exists in non-human viruses ($R = 0.08$, $p = 0.44$). Thus, through multiple measures within and between viruses, one can clearly see a correlation between the *SIR* score and the protein length.

To summarize, these results suggest that the relation between the organism's complexity and the fixation of advantageous mutations extends from the single protein to the full organism. The difference between viruses infecting humans and viruses infecting other species suggests that these results are indeed due to selection against epitope presentation on human HLA molecules and not to properties of short and long proteins in general or generic features of small and large viruses. One could argue that this effect

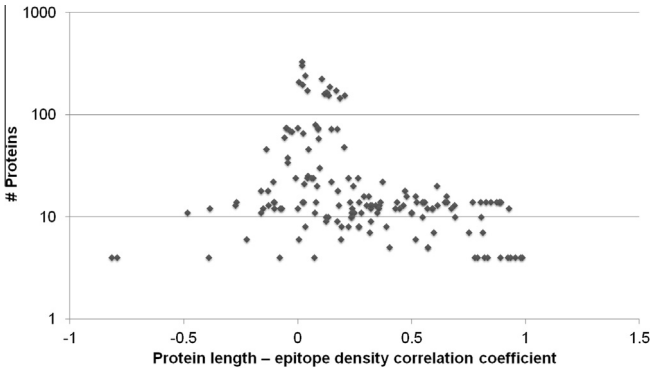

Fig. 4. Number of proteins in virus versus the correlation coefficient of *SIR* score and protein length. The *x* axis is the coefficient of the SIR score – protein length correlation and the y axis is the number of proteins. Viruses with a small number of proteins have a higher *SIR* score-protein length correlation than viruses with large number of proteins. ($R = -0.27$, $p < 1.e{-}5$). This can be very clearly seen by the large number of small viruses with high correlation coefficients (lower right part of the distribution), and the absence of a parallel distribution of small viruses with negative correlation coefficients. Epitope prediction was done by the MLVO algorithm.

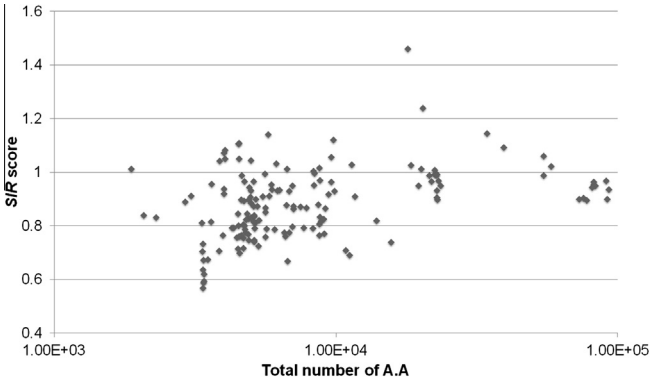

Fig. 5. Virus proteome size versus its averaged *SIR* score. For each virus, the lengths of all of its proteins were summed and tested for correlation with the *SIR* score. The *x* axis is the total number of A. A in a virus (the sum of the protein lengths in the virus), and the *y* axis is the average *SIR* score for the same virus. One can see that this correlation is stronger than the correlation of the *SIR* score with either protein length or number of proteins alone (Spearman $R = 0.41$, $p < 1.e{-}10$). Epitope prediction was done by the MLVO algorithm.

is due to the similarity at the sequence level between proteins in the group of viruses. However, even when all viruses belonging to the same group are clustered to a single point, the relation between the number of proteins and the average *SIR* score can be clearly observed.

Given the correlation between the *SIR* score, and both the protein length and their number, we hypothesized that their combination is even more correlated with the *SIR* score. To test that, we computed the correlation of the *SIR* score with the total proteome size for each virus (the sum of the virus protein lengths). As expected, this combination resulted in an even higher correlation ($R = 0.41$, $p < 1.e{-}10$) (Fig. 5).

3. Discussion

We have here shown, using bioinformatics tools and large scale genetic data sets that selection for escape mutations (specifically, mutations that remove CTL epitopes) in viruses is mainly focused on short proteins and small viruses. Such results are expected given that in large proteins/viruses the removal of epitopes has a fitness cost, but no significant survival advantage. This is an extension of previous models on the incompatibility between the fitness cost and the phenotypic advantage of each mutation in complex organisms (Orr, 2000; Wagner and Altenberg, 1996). As the organism becomes more complex, the probability that a mutation should increase the organism's fitness decreases, while the cost of each mutation stays constant.

A positive correlation was here described between the epitope density (as measured by the *SIR* score using the MLVO MHC-I binding prediction algorithm) of each protein and the protein length. A similar correlation was observed between the average epitope density in a full virus and the number of proteins in the virus. The results were also validated using the classical, yet less precise, BIMAS algorithm for MHC binding (Fig. S5, Table S1). These correlations were observed in viruses infecting humans, and to a much lesser extent in viruses infecting non-human hosts.

It is important to mention that the proteasome and the TAP channels which are highly conserved among species and human MHC also show some level of similarity to non-human MHC, and hence a correlation was seen in non-human viruses as well. However, the stronger correlation in human viruses shows that the reduction in the epitope number is indeed the result of immune-induced selection against epitope presentation.

In order to avoid the destruction of its host cell, the virus evolves to remove a large fraction of the epitopes. Removing a limited part of the epitopes has a very limited advantage and can have a high fitness cost. Thus, even if the cost per mutation is larger in small viruses (and the more so if it is constant), the increase in the number of total required mutations makes it harder for large viruses to adapt. Thus, the very clear negative correlation between the number of proteins and the accumulation of escape mutations may be a result of the "all or none" selection force affecting viruses.

An alternative explanation for the negative correlation between selection against epitopes and the number of proteins is that viruses with a large number of proteins have a higher probability of expressing immune regulatory proteins and hence are less threatened by CTL recognition. However, it seems that this is not the main factor that determines selection against epitopes, since some of the small viruses do express immune-regulatory proteins (HIV Vigerust et al., 2005; Piguet and Trono, 2001; Piguet et al., 2004, HCV (Zimmermann et al., 2008; Kim et al., 2012), and they, as other small viruses, have a low epitope density.

Among all of the peptides that presented on the MHC-I molecule, only a small fraction will eventually induce T cell response (i.e. Immunodominant epitopes (Yewdell, 2006).

Although most presented peptides will probably not induce a T cell response, the systematic removal of these peptides will lower the probability of appearance of immunodominant epitopes. Therefore a lower number of presented peptides may account for a stronger selection against T cell response.

Note that if indeed a small number of presented peptides are immunodominant, and many of the peptides computed to be presented in the current analysis do not induce a T cell response, then the observed decrease of the average epitope density to 70% of its expected value in small viruses, may actually represent a removal of all immunodominant epitopes. In such a case, the effect of the viral complexity may actually be much more significant that we present here.

From a theoretical point of view, immunodominance is also expected to increase the effect of the viral complexity on the accumulation of escape mutations. Assume that only very few presented peptides can induce a strong immune response. In such a case, for a small virus it is enough to mutate a few epitopes and completely escape from the immune system. For large viruses, it may be impossible to remove enough epitopes to prevent an immunodominant response, and as such gain nothing from escape mutations.

Although epitope density is used as a measure for selection, the relation between these two is not straight forward. In previous studies we have demonstrated that inherent characters of a protein influence its epitope density regardless to the presence of selection (Maman et al., 2011a). For example, hydrophobic proteins naturally have more epitopes that hydrophilic proteins, due to the nature of the MHC-I binding groove. Therefore, viruses that have an hydrophobic proteins, might have higher SIR score even though they are under strong selection for epitope removal. One striking example for this is the Human coronavirus that have low number of proteins but relatively high average SIR score (1.46) (Fig. 2, Table S1).

More generally, we have shown in the past that Viruses infecting humans have less epitopes that viruses infecting non-human hosts on human HLA alleles. Within the human viruses, there are multiple factors affecting the epitope density. We have here shown that the complexity of the virus is one of the major elements shaping this density, but not the only one.

The genetic complexity as represented by the number of proteins and their length does not completely capture the phenotypic complexity, which is much more complex to define and measure. However, the presence of such a clear correlation suggests that these measures are at least related to the organism complexity.

Most recent studies on evolution have been done in a regime of weak selection and low mutation rate (Rokyta et al., 2005, 2006; Fudenberg et al., 2006; Lande, 2009) (for a review see Orr, 2005). One could have assumed that in viruses, with the extreme regime of high levels of both selection and mutation rates, the relation between complexity and adaptation would be lost, and viruses would be able to optimally adapt their sequence to avoid detection. We have here shown that even in such extreme cases, a balance between complexity and adaptation exists.

These results have implications far beyond the specific issue of escape mutations. We have shown that beyond 40 proteins, viruses fail to adapt their genome to the host immune system. Actually, beyond 40 proteins, there is practically no adaptation. One can therefore ask, how can much more complex organisms, with a much lower mutation rate (1.e−4 vs 1.e−9), a much longer life-cycle (hours vs. years), and a much smaller population (thousands to billions per species for most advanced species vs. more than 1.e10 in each different host for many viruses) evolve to adapt to their environment.

The simple answer may be modularity. We have previously shown in the case of herpesviruses and bacteria (Vider-Shalit et al., 2007; Maman et al., 2011c) that while most proteins do not avoid detection, limited groups of proteins, such as Herpesvirus latent protein, or Type III secretion system effectors of gram-negative bacteria do accumulate escape mutations (Vider-Shalit et al., 2007; Maman et al., 2011c). The same thing may be true for the evolution of advanced species: while the full genome (or even groups of tens to hundred ore genes) is way too complex to adapt, limited gene groups may adapt to their environment.

4. Material and methods

4.1. SIR score

We have analyzed the ratio between the number of epitopes presented in viral proteins and the number of epitopes in random proteins with the same length and typical viral amino acid composition. This ratio was defined as the Size of Immune Repertoire (*SIR*) score. The epitope number was computed using three algorithms: a proteasomal cleavage algorithm (Ginodi et al., 2008), a TAP binding algorithm developed by Peters et al. and the MLVO MHC binding (Vider-Shalit and Louzoun, 2010) algorithms. The algorithms' quality was systematically validated using epitope databases and was found to induce low FP and FN error rates. Different alleles present different set of epitopes. Thus, the analysis is first performed at the single allele level. For instance, if a sequence from a viral protein X has 4 epitopes that can bind the groove of the HLA allele A∗0201 and a random sequence with a similar length and a typical viral amino acid distribution is expected to have 10 HLA A∗0201 epitopes, then the *SIR* score of X for HLA A∗0201 would be 0.4 (4/10). We have computed epitopes for the 39 most common HLA alleles and weighted the results according to the allele frequency in the Caucasian population (Newell et al., 1996). The computation of the *SIR* scores can be performed through our web-server at http://peptibase.cs.biu.ac.il/index.html.

4.2. Cleavage score

Given a peptide with N- and C-terminal flanking residues FN and FC and residues

P_1,. .P_i,.. P_n, where P_i represents any residue 1, and n represents C and N positions, the following score was defined:

$$S(\text{peptide}) = S_1(FN) + S_2(P_1) + \sum_{i=2}^{n-1} S_3(P_i) + S_4(P_n) + S_5(FC).$$

A peptide with a high score, *S*, has a high probability of being produced, while a low score corresponds to a low probability of production. The appropriate values for S_1 to S_5 were learned using a simulated annealing process. The algorithm was validated to give a rate of false positives of less than 16% and a rate of false negatives of less than 10% (Ginodi et al., 2008).

4.3. MHC binding analysis using multi-label vector optimization (MLVO)

The MLVO algorithm (Vider-Shalit and Louzoun, 2010) for MHC binding prediction finds a classifier (w) using three label types that are combined into a single constrained optimization problem. The method finds the optimal combination of binary classification of peptides known to bind or not to bind the MHC molecule, a linear regression based on the measured affinities of peptides with a known IC50 or EC50 binding concentrations and a guess (often based on information on similar alleles). Solving this optimization problem results in a Position Weight Matrix for each HLA alleles. These matrices estimate the contribution of each amino acid at each position to the total binding strength. The accuracy of MHC binding prediction for the vast majority of MHC-I alleles in the MLVO is over 0.95 (with AUC of over 0.98). As in all other cases, the *SIR* results presented are an average weighted over alleles of the ratio between the computed epitope density and the one expected in a random sequence. The *SIR* scores of the viral proteins in this study are presented in Supplementary Material Table S1.

4.3. Thresholds

The MHC binding prediction algorithm provides a binding score for each nine-mer. In order to produce an epitope list, a cutoff should be applied to these scores for each allele. The way the cutoff is determined is based on the competition for the presentation on a limited number of MHC molecules. For example, an allele such as B∗2705 is expected to present a very large number of epitopes from self proteins. Thus a viral protein with a large number of epitopes would have to compete with a similarly high number of epitopes in human proteins. While this approach may lead to the exclusion of some real viral epitopes, it should not affect the ratio between the number of computed epitopes in human and non-human viruses.

4.4. Epitope computation server

We have designed a CTL epitope SQL based library webserver (http://peptibase.cs.biu.ac.il). This website provides detailed CTL epitope libraries for the human and mouse genomes as well as for most fully sequenced viruses. It also allows users to upload a file and produce an epitope library. All viral proteins in this study were analyzed for their epitope using this webserver.

4.5. Statistics

All comparisons were performed using two-sided unequal variance T tests. The correlation between length and *SIR* score was computed using a Spearman Correlation since the distribution of the protein lengths is approximately log normal (Supplementary Material Fig. S6) and not normal.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2013.07.030.

References

Aebischer, T., Moskophidis, D., Rohrer, U.H., Zinkernagel, R.M., Hengartner, H., 1991. In vitro selection of lymphocytic choriomeningitis virus escape mutants by cytotoxic T lymphocytes. Proc. Natl. Acad. Sci. USA 88 (24), 11047–11051.

Agranovich, A., Vider-Shalit, T., Louzoun, Y., 2011. Optimal viral immune surveillance evasion strategies. Theor. Popul. Biol. 80 (4), 233–243, PubMed PMID: 21925527. Pubmed Central PMCID: 3218222. Epub 2011/09/20. eng.

Alcami, A., 2003. Viral mimicry of cytokines, chemokines and their receptors. Nat. Rev. Immunol. 3 (1), 36–50.

Baatz, M., Wagner, G.P., 1997. Adaptive inertia caused by hidden pleiotropic effects∗ 1. Theor. Popul. Biol. 51 (1), 49–66.

Bowen, D.G., Walker, C.M., 2005 Jun 6. Mutational escape from CD8+ T cell immunity: HCV evolution, from chimpanzees to man. J. Exp. Med. 201 (11), 1709–1714.

Delport, W., Scheffler, K., Seoighe, C., 2008. Frequent toggling between alternative amino acids is driven by selection in HIV-1. PLoS Pathog. 4 (12), e1000242, PubMed PMID: 19096508. Pubmed Central PMCID: 2592544. Epub 2008/12/20. eng.

Fisher, R.A., 1930. The Genetical Theory Of Natural Selection. Oxford University Press, Oxford, UK.

Fudenberg, D., Nowak, M.A., Taylor, C., Imhof, L.A., 2006. Evolutionary game dynamics in finite populations with strong selection and weak mutation. Theor. Popul. Biol. 70 (3), 352–363.

Gillespie, J.H., 1983. A simple stochastic gene substitution model. Theor. Popul. Biol. 23 (2), 202–215.

Gillespie, J.H., 1984. Molecular evolution over the mutational landscape. Evolution, 1116–1129.

Gillespie, J.H., 1994. The causes of molecular evolution. Oxford University Press, USA.

Ginodi, I., Vider-Shalit, T., Tsaban, L., Louzoun, Y., 2008. Precise score for the prediction of peptides cleaved by the proteasome. Bioinformatics 24 (4), 477–483, PubMed PMID: 18216070. Epub 2008/01/25. eng.

Kim, H., Mazumdar, B., Bose, S.K., Meyer, K., Di Bisceglie, A.M., Hoft, D.F., et al., 2012. Hepatitis C virus-mediated inhibition of cathepsin S increases invariant-chain expression on hepatocyte surface. J. Virol. 86 (18), 9919–9928, PubMed PMID: 22761382. Pubmed Central PMCID: 3446550. Epub 2012/07/05. eng.

Kimura, M., 1983. The neutral theory of molecular evolution. Cambrige University Press.

Kovjazin, R., Ilan, V., Yair, D., Vider-Shalit, T., Azran, R., Tsaban, L., et al., 2011. Signal peptides and trans-membrane regions are broadly immunogenic and have high CD8+ T cell epitope densities: Implications for vaccine development. Mol. Immunol., In Press. PubMed PMID: 19730693. Pubmed Central PMCID: 2731216. eng.

Lande, R., 2009. The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genet. Res. 26 (03), 221–235.

Macken, C.A., Perelson, A.S., 1984. A multistage model for the action of cytotoxic T lymphocytes in multicellular conjugates. J. Immunol. 132 (4), 1614.

Maman, Y., Blancher, A., Benichou, J., Yablonka, A., Efroni, S., Louzoun, Y., 2011a. Immune induced evolutionary selection focused on a single reading frame in overlapping HBV proteins. J. Virol., JVI. 02142-10v1.

Maman, Y., Nir-Paz, R., Louzoun, Y., 2011b. Bacteria modulate the CD8+ T cell epitope repertoire of host cytosol-exposed proteins to manipulate the host immune response. PLoS Comput. Biol. 7 (10), e1002220.

Maman, Y., Nir-Paz, R., Louzoun, Y., 2011c. Bacteria modulate the CD8+ T cell epitope repertoire of host cytosol-exposed proteins to manipulate the host immune response. PLoS Comput. Biol. 7 (10), e1002220, PubMed PMID: 22022257. Pubmed Central PMCID: 3192822. Epub 2011/10/25. eng.

McMichael, A.J., Phillips, R.E., 1997. Escape of human immunodeficiency virus from immune control. Annu. Rev. Immunol. 15, 271–296.

Newell, W.R., Trowsdale, J., Beck, S., 1996. MHCDB: database of the human MHC (release 2). Immunogenetics 45 (1), 6–8.

Orr, H.A., 2000. Adaptation and the cost of complexity. Evolution 54 (1), 13–20.

Orr, H.A., 2002. The population genetics of adaptation: the adaptation of DNA sequences. Evolution 56 (7), 1317–1330.

Orr, H.A., 2003. The distribution of fitness effects among beneficial mutations. Genetics 163 (4), 1519.

Orr, H.A., 2005. The genetic theory of adaptation: a brief history. Nat. Rev. Genet. 6 (2), 119–127.

Piguet, V., Trono, D., 2001. Living in oblivion: HIV immune evasion. Semin. Immunol. 13 (1), 51–57, PubMed PMID: 11289799. Epub 2001/04/06. eng.

Piguet, V., Wan, L., Borel, C., Mangasarian, A., Demaurex, N., Thomas, G., et al., 2004. HIV-1 Nef protein binds to the cellular protein PACS-1 to downregulate class I major histocompatibility complexes. Nat. Cell Biol. 2 (3), 163–167, PubMed PMID: 10707087. Pubmed Central PMCID: 1475706. Epub 2000/03/09. eng.

Poppema, S., Potters, M., Visser, L., Van Den Berg, A.M., 1998. Immune escape mechanisms in Hodgkin's disease. Ann. Oncol. 9 (Suppl. 5), S21.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35 (Database issue), D61–D65, PubMed PMID: 17130148. Pubmed Central PMCID: 1716718. Epub 2006/11/30. eng.

Regoes, R.R., Barber, D.L., Ahmed, R., Antia, R., 2007. Estimation of the rate of killing by cytotoxic T lymphocytes in vivo. Proc. Natl. Acad. Sci. 104 (5), 1599.

Regoes, R.R., Yates, A., Antia, R., 2007. Mathematical models of cytotoxic T-lymphocyte killing. Immunol. Cell Biol. 85 (4), 274–279.

Rokyta, D.R., Joyce, P., Caudle, S.B., Wichman, H.A., 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat. Genet. 37 (4), 441–444.

Rokyta, D.R., Beisel, C.J., Joyce, P., 2006. Properties of adaptive walks on uncorrelated landscapes under strong selection and weak mutation. J. Theor. Biol. 243 (1), 114–120.

Sanjuan, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral Mutation Rates. J. Virol. 84 (19), 9733–9748.

Soderholm, J., Ahlen, G., Kaul, A., Frelin, L., Alheim, M., Barnfield, C., et al., 2006. Relation between viral fitness and immune escape within the hepatitis C virus protease. Gut 55 (2), 266–274.

Timm, J., Lauer, G.M., Kavanagh, D.G., Sheridan, I., Kim, A.Y., Lucas, M., et al., 2004. CD8 epitope escape and reversion in acute HCV infection. J. Exp. Med. 200 (12), 1593.

Vider-Shalit, T., Louzoun, Y., 2010. MHC-I prediction using a combination of T cell epitopes and MHC-I binding peptides. J. Immunol. Meth., PubMed PMID: 20920507. Pubmed Central PMCID: 3044214. Epub 2010/10/06. Eng.

Vider-Shalit, T., Fishbain, V., Raffaeli, S., Louzoun, Y., 2007. Phase-dependent immune evasion of herpesviruses. J. Virol. 81 (17), 9536–9545, PubMed PMID: 17609281. Pubmed Central PMCID: 1951411. Epub 2007/07/05. eng.

Vider-Shalit, T., Almani, M., Sarid, R., Louzoun, Y., 2009a. The HIV hide and seek game: an immunogenomic analysis of the HIV epitope repertoire. AIDS 23 (11), 1311–1318, PubMed PMID: 19550286. Epub 2009/06/25. eng.

Vider-Shalit, T., Sarid, R., Maman, K., Tsaban, L., Levi, R., Louzoun, Y., 2009b. Viruses selectively mutate their CD8+ T-cell epitopes-a large-scale immunomic analysis. Bioinformatics 25 (12), i39–i44.

Vigerust, D.J., Egan, B.S., Shepherd, V.L., 2005. HIV-1 Nef mediates post-translational down-regulation and redistribution of the mannose receptor. J. Leukoc. Biol. 77 (4), 522–534, PubMed PMID: 15637102. Epub 2005/01/08. eng.

Wagner, G.P., 1996. Homologues, natural kinds and the evolution of modularity. Integr. Comp. Biol. 36 (1), 36.

Wagner, G.P., Altenberg, L., 1996. Perspective: complex adaptations and the evolution of evolvability. Evolution 50 (3), 967–976.

Welch, J.J., Waxman, D., 2003. Modularity and the cost of complexity. Evolution 57 (8), 1723–1734.

Williams, A., Peh, C.A., Elliott, T., 2002. The cell biology of MHC class I antigen presentation. Tissue Antigens 59 (1), 3–17, PubMed PMID: 11972873. Epub 2002/04/26. eng.

Yates, A., Graw, F., Barber, D.L., Ahmed, R., Regoes, R.R., Antia, R., 2007. Revisiting estimates of CTL killing rates in vivo. PLoS One 2 (12), 1301.

Yewdell, J.W., 2006. Confronting complexity: real-world immunodominance in antiviral CD8+ T cell responses. Immunity 25 (4), 533–543, PubMed PMID: 17046682. Epub 2006/10/19. eng.

Zimmermann, M., Flechsig, C., La Monica, N., Tripodi, M., Adler, G., Dikopoulos, N., 2008. Hepatitis C virus core protein impairs in vitro priming of specific T cell responses by dendritic cells and hepatocytes. J. Hepatol. 48 (1), 51–60, PubMed PMID: 17998148. Epub 2007/11/14. eng.