Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Immunity space generated by a non trivial genetic-antigenic relation

by

Lorenzo Taggi, Francesca Colaiori, Vittorio Loreto, and Francesca Tria

Preprint no.: 27 2012

Immunity space generated by a non trivial genetic–antigenic relation

Lorenzo Taggi[1], Francesca Colaiori[3], Vittorio Loreto[2,4] and Francesca Tria[4]
[1] *Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103, Leipzig, Germany*
[2] *Sapienza University of Rome, Physics Department, P.le A. Moro 5, 00185 Rome, Italy*
[3] *CNR-ISC, P.le A. Moro 5, 00185 Rome, Italy and*
[4] *ISI Foundation, Viale Settimio Severo 65, 10133, Torino, Italy*
(Dated: September 2011)

Cross–immunity plays a crucial role in the epidemiological properties of influenza A virus. Recent experimental studies have pointed out that genetic distances do not fully account for the observed antigenic clusters. In particular, jumps from an antigenic cluster to another seem to be triggered by correlated mutations. However, no specific sites determine whether a sequence belongs to a given antigenic cluster, suggesting that sites mutations could be dynamically correlated. In this paper we introduce an epistatic interaction rule among mutations that dynamically defines neutral clusters in the immunity space. We investigate the structure of this epistatic immunity space, highlighting how this can affect the dynamical properties of the virus–host interaction.

The interest of the scientific community in the Influenza A virus evolution has been continuously increasing in the last years [1–3]. Understanding the mechanisms that drive the ever–changing of the antigenic determinants is crucial in order to implement effective prevention strategies. Major efforts have been devoted to explain apparently contradictory features: On one hand the virus mutates fast enough so that the same host can be infected several times in the course of its life, on the other hand a viral quasispecies can be sufficiently well defined in any given epidemic season, so that a temporarily effective vaccine can be developed. The peculiar evolutionary dynamics of the Influenza A virus is revealed by the comb–like shape of its phylogenetic tree [4–6], as reconstructed from haemaglutinin (HA) coding sequences. It has been contrasted with phylogenetic trees of other viruses [7], as measles virus and HIV virus at the population level, which show more ramified patterns [8].

A crucial mechanism driving the interaction between the virus and the host immune system is *cross–immunity*: After being infected by a strain, the host acquires partial or total immunity to a set of other strains *antigenically similar* to the infecting one [9]. However it is not yet clear what determines the similarity relation in terms of genetic distance. A first attempt to reproduce in a modeling framework the complex balance between strains proliferation induced by antigenic drift, and strains selection, induced by the increasing acquired immunity of the hosts, is due to Ferguson et al. [1]. In their work, a mechanism of broad spectrum cross immunity, lasting for a period of several weeks after infection, in addition to the life–long cross–immunity, is claimed to be crucial in order to recover the observed evolutionary dynamics of the Influenza A virus. Although this idea seems to be confirmed in the framework of simple evolutionary models [10, 11], a clear evidence of the existence of such a mechanism has not been provided so far.

A common trait of the above mentioned and previous models [12] is the assumed equivalence between genetic and antigenic distance: mutations in the HA protein accumulate in time until eventually the mutated strain becomes enough antigenically distant to escape host immunity. In this case the degree of cross–immunity between the two strains is measured in terms of the Hamming distance between their sequences. Recent studies, however, highlight how that assumption is not completely correct [13]: high genetic differences can be irrelevant from the antigenic point of view and, vice versa, few nucleotidic mutations can elicit a large antigenic effect [13, 14], indicating that the accumulation of genetic distance is not a necessary (and sometimes nor sufficient) condition for the emergence of antigenically novel strains. Moreover, genetic and antigenic evolution exhibit different patterns: While genetic mutations occur gradually, antigenic evolution seems more punctuated, though a debate is still going on [15]. Further, it has been pointed out that amino acid changes which seem to be relevant in differentiating two specific antigenic clusters, can exhibit a null antigenic effect when appearing in different sequences [13], suggesting that antigenic clusters cannot simply be associated with key influential sites [16]. The presence of correlations in genetic mutations might explain why phenotypic changes do not necessarily appear as a consequence of accumulated mutations. Correlation between mutations have indeed been observed [15] and positive epistasis between pairs of sites in neuraminidase (NA) and hemagglutinine (HA) proteins is supported by phylogenetic and sequence analisys [17–19]. The effect on the Influenza virus evolutionary dynamics of a non-trivial relation between genotypic and phenotypic (antigenic) space has been investigated introducing a neutral network topology in the space of sequences [20, 21]. We here investigate the effect of correlations between point mutations. In particular, we consider correlations between pairs of sites, though the present framework can be easily extended to more complex patterns of correlated mutations. Our aim is to investigate the geometric properties of the resulting antigenic space and to possibly

relate them to the viral evolutionary dynamics. We find that the introduction of dynamic correlations reflects in a staggered time structure, with an alternation of periods where a high number of relatively low fitness strains are able to spread the infection, followed by periods where a single highly fit strain is the favoured escape mutant. This behaviour is absent when we consider the antigenic distance as directly proportional to the genetic distance.

The Epistatic Immunity Space. We represent viral strains by binary sequences $\vec{v}$ of fixed length n [22]. We define the immunity set $I_n(\vec{v})$ of a strain $\vec{v}$ as the set of viruses antigenically similar to it: those viruses that cannot infect an host that has been already infected by $\vec{v}$. We can further consider the immunity elicited by more than one strain, for istance by all the strains produced by successive mutations and spread during an infection history. We call the *Epistatic Immunity Space* (EIS) $I_n(A)$ of the *infection set* A the union of all the immunity sets $I_n(\vec{v})$ of the strains in A:

$$I_n(A) = \bigcup_{\vec{v}\in A} I_n(\vec{v}). \tag{1}$$

The immunity set, and therefore the EIS, depends on the definition of *antigenic similarity*. We investigate the simplest choice which includes correlations: We assume that two strains are cross–immune unless they differ in at least two consecutive bits [23]. Therefore

$$I_n(\vec{v}) = \{\, \vec{z} \in H_n :\; z_i \neq v_i \Rightarrow\; z_{|i+1|_n} = v_{|i+1|_n}\; \forall i \,\}, \tag{2}$$

where H_n is the n–dimensional hypercube, composed of 2^n strings, with the metric given by the Hamming distance, and periodic boundary conditions. The fraction $\rho_n(i)$ of strains that belong to $I_n(\vec{v})$ and have Hamming distance i from $\vec{v}$ can be computed (see SI) and reads $\rho_n(i) = \exp(-i^2/n) + O(1/n)$: correlations introduce on the one hand a non trivial correspondance between genotipic and phenotipic space, on the other hand antigenic similarity is not completely decorrelated from genetic distance [13]. The size $S(n) \equiv |I_n(\vec{v})|$ of the immunity set generated by a strain, i.e. the number of strains cross–immune to it, satisfies a Fibonacci–like recursive relation: $S(n) = S(n-1) + S(n-2)$ with initial condition $S(2) = 3$ and $S(3) = 4$. $S(n)$ is known as Lucas sequence, and an explicit expression is known: $S(n) = \phi^n + (1-\phi)^n \simeq \phi^n$, where $\phi = (1+\sqrt{5})/2 \sim 1.618$ is the golden ratio, and the last asymptotic holds for large n (see also SI). The size $|I_n(A)|$ of the *Epistatic Immunity Space* (EIS) generated by k different strains strongly depends on the actual form of the set A. We firstly consider two quantities that provide bounds for every epidemic dynamics with the above defined antigenic similarity measure: (*i*) $M(n)$, the maximum number of distinct strings that fit in the sequence space, and such that the next string would immunize the whole space (the strings are therefore chosen with the maximum overlap between their immunity sets); (*ii*) $m(n)$, the minimum number of strings needed to immunize the whole sequence space, and therefore chosen with the minimum overlap between their immunity sets. The computation of $M(n)$ is straightforward: in order to have at least a string, say $\vec{v}$, left out of the EIS, the infection set cannot contain *any* of the strings in $I(\vec{v})$. Therefore, the biggest infection set that does not immunize the whole hypercube is $A_{D(n)} = H_n \,/\, I(\vec{v})$, which immunizes the set $H_n/\{\vec{v}\}$, and $M(n) = 2^n - S(n)$. We estimate $m(n)$ by numerical simulations and we provide analytically an upper $m_U(n)$ and a lower $m_L(n)$ bound. A (trivial) lower bound is given by assuming totally disjoint immunity sets, and it is given by counting the total number of sequences divided by the size $S(n)$ of a single immunity set: $m_L(n) \simeq 2^n/\phi^n = 2^{\eta n}$, with $\eta = 1 - \ln_2 \phi \sim 0.306$. The fraction of strings contained in the immunity set of a single strain is therefore $2^{-\eta n}$. An upper bound $m_U(n)$ can be derived constructively by exhibiting a set of sequences whose immunity sets cover the sequence space. Such a set of sequences is obtained for example by combining in all possible ways $n/2$ pairs of identical bits, either $(0,0)$ or $(1,1)$ (for instance for $n = 4$ such a coverage is realized by the four sequences $(0,0,0,0), (1,1,0,0), (0,0,1,1), (1,1,1,1)$). The number of such sequences is $m_U(n) = 2^{\left[\frac{n}{2}\right]}$, where $[\cdot]$ denotes the integer part. The asymptotic value of $m(n)$ as computed numerically by simulated annealing [24] (see SI) is $m(n) \simeq 2^{\alpha n}$ with $\alpha \sim 0.4$, compatible with the analytical bounds.

Let us now focus on the topological properties of the EIS. Noticeably, the EIS is always a connected set, for any infection history. This can be seen as follows: we need to show that for any pair of sequences $\vec{x}$, $\vec{y}$, there exists a path of cross-immune sequences joining them. It is easy to convince oneself that every single immunity set is connected. It is thus enough to show that any pair of immunity sets overlap, or are at most contiguous. Take $\vec{x} = \vec{0}$ without loss of generality, and $\vec{y} = (y_1, y_2, \ldots, y_n)$. For n even, the two immune sets always overlap at least in the sequence $(0, y_2, 0, y_4, 0, \ldots, 0, y_n)$. For n odd they are contiguous in the two points $(0, y_2, 0, y_4, 0, \ldots, 0, y_{n-1}, 0) \in I_n(\vec{0})$ and $(0, y_2, 0, y_4, 0, \ldots, 0, y_{n-1}, y_n) \in I_n(\vec{y})$ (actually they always overlap at some point unless $\vec{y} = \vec{1}$).

Though always connected, the EIS is not always simply connected, and the complementary set, i.e. the infectious region, can be not connected. This might have a strong impact on the underlying virus–host interaction. For example, the EIS for k strings drawn at random is simply connected only for $k = \lceil n/2 \rceil$ (see SI). For k slightly above this threshold the infectious region is composed by one big connected cluster and many small connected clusters ("holes" in the EIS). The disappearence of the big connected cluster as k increases sets the threshold where a further spread of an epidemic is inhibited (see

Fig. 1 for a cartoon and SI for a more detailed analysis).

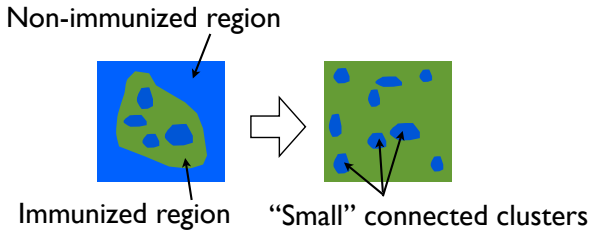

FIG. 1. Sketch of the noninfectious (green) and infectious (blue) region of the sequence space. Left: for small k above threshold the infectious region has big connected cluster, corresponding to a infectious region of the hypercube, and small connected clusters. Right: increasing k only small holes in the EIS are left.

We have so far examined general topological properties of the EIS from a static point of view. The nontrivial shape of the set of cross–immune sequences can be however better highlighted considering simple infectious dynamics. We then consider a local maximization (LM) of the EIS: starting with a random strain, we chose at every step the next strain among those not already belonging to the EIS, and such that it maximizes the size of the current EIS (and thus minimizes the overlap with the existing EIS). In case that several strings satisfy this criterion we chose one at random among them. We iterate until the whole space H_n is noninfectious. Note the this dynamics performs a *local* maximization, therefore we expect the number of strains required to immunize the whole space to be greater than $m(n)$. Though in a very naïve way, the local maximization of the epistatic immunity set mimicks a successful escape strategy for the virus in an actual population: smaller is the overlap of the new strain immunity set with the pre–existing EIS, smaller would be the probability for the host to be already immune to the new strain.

If we look at the number of sequences that satisfy the local maximization constraint at each time step, we find a peculiar behaviour which is not observed when the immunity sets are constructed by means of the bare hamming distance from the generating strain. The time behavior features a well defined series of peaks corresponding to an alternation of periods with many equivalent options and only one optimal option to maximize the immunity set (Fig. 2): this gives a hint of how dynamical constraints arise from the presence of epistatic interactions with respect to the case in which antigenic distance is directly proportional to genetic distance.

To further characterize the epidemic dynamics we look at the *normalized invasion rate*, i.e. the fraction of strains becoming noninfectious at each step of the LM dynamics (Fig. 3). This quantity also shows a non-trivial behavior characterized by a series of hierarchically distributed jumps that occur always at the same time steps, independently of n, and that are not present when the

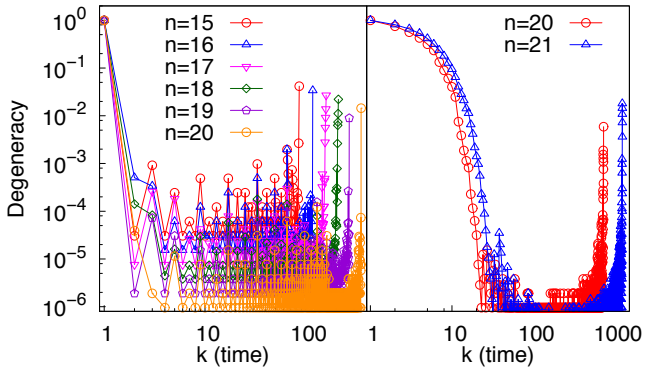

FIG. 2. Left: Degeneracy of strings allowed by the local optimization dynamics (normalized with 2^n) as a function of the iteration number k (time) for the epistatic rule. The averages are taken over 1000 realizations. Right: For comparison: same dynamics, but with cross–immunity defined by the Hamming rule with distance $D = 4$ (the immunity set is the set of all strings whose Hamming distance $\leq D$ from the generating one).

same dynamics is studied with a Hamming rule for cross–immunity (see Fig. 3, right). This points again to a staggered time structure with an alternation of periods of highly effective immunization, followed by periods with a relatively lower immunization rate. This picture is also confirmed by the parametric plot in the bottom of Fig. 3 where the degeneracy (the fraction of optimal strains) is plotted versus the normalized invasion rate. The peculiar triangular structure, absent in the Hamming case, is the signature of an alternation of times with no degeneracy (only one option) corresponding to a high invasion rate followed by times with a very high degeneracy and low invasion rate. This behaviour is reminiscent of the comb-like shape of the Influenza HA phylogenetic tree, where a single quasispecies is responsible for each annual epidemic and antigenic clusters follow one another each few years [14].

Conclusion and perspectives. In this paper we introduced a specific interaction rule among mutations that dynamically defines neutral clusters in the immunity space. We have studied how this rule shapes the EIS, i.e. the set of viruses to which a host is immune after infection by all the strains in his infection history. We have studied in particular a simple extremal dynamics of viral epidemic, focusing on the important differences with respect to the case where the usual Hamming distance defines cross–immunity. A striking difference that we find is represented by a staggered time structure, in the epistatic case, in which times where one single choice exists that maximizes the invasion rate are followed by times where many different options exist to immunize a relatively smaller set of sequences. Although our model is a toy model, if one imagines this feature in a realistic virus–host dynamics, it is quite tempting to iden-

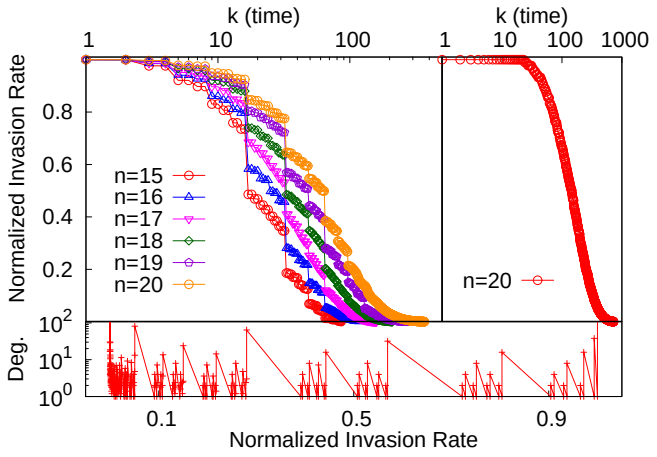

FIG. 3. Top left: Time behavior of the normalized invasion rate, i.e., the fraction of sequences becoming noninfectious at time k for different values of n. Top right: For comparison: Same quantity but with cross–immunity defined by the Hamming rule. In this case no jumps are observed. Bottom: Parametric plot of the Degeneracy vs. the Normalized Invasion Rate for the epistatic rule with $n = 19$.

tify our staggered structure with the succession in time of different antigenic clusters and with the more violent epidemic outbreaks at each cluster change. The analysis presented here can help in understanding the effect of the conjectured epistatic interactions on the shape of immunity clusters as well as on the viral evolutionary dynamics at large, with a possibly relevant impact on the structure of the modeling schemes. The analysis of the effect of the epistatic rule in a more realistic model of virus evolution in a host population is the subject of a forthcoming work.

[1] N. M. Ferguson, A. P. Galvani, and R. M. Bush, Nature **422**, 428 (2003).

[2] M. Nelson and E. Holmes, Nat. Rev. Gen. **8**, 196 (2007).

[3] D. Vijaykrishna, G. J. D. Smith, O. G. Pybus, H. Zhu, S. Bhatt, L. L. M. Poon, S. Riley, J. Bahl, S. K. Ma, C. L. Cheung, et al., Nature **473**, 519 (2011), ISSN 0028-0836, URL `http://dx.doi.org/10.1038/nature10004`.

[4] W. Fitch, R. Bush, C. Bender, and N. Cox, Proc. Natl. Acad. Sci. USA (PNAS) **94**, 7712 (1997).

[5] R. Bush, C. Bender, K. Subbar, N. Cox, and W. Fitch, Science **286**, 1921 (1999).

[6] R. Bush, C. Smith, N. Cox, and W. Fitch, Proc. Natl. Acad. Sci. USA (PNAS) **97**, 6974 (2000).

[7] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes, Science **303**, 428 (2003).

[8] The phylogenetic tree of HIV virus inside a single host, where selective pressure plays a crucial role, presents instead similar features as the Influenza A virus tree.

[9] P. W. Gill and A. M. Murphy, Med. J. Aust. **2**, 761 (1977).

[10] F. Tria, M. Lassig, L. Peliti, and S. Franz, J. Stat. Mech. p. P07008 (2005).

[11] G. Bianconi, D. Fichera, S. Franz, and L. Peliti (2010), URL `http://arxiv.org/abs/1010.3539`.

[12] M. Girvan, D. Callaway, M. Newman, , and S. Strogatz, Phys. Rev. E p. 031915 (2002).

[13] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. A. M. Fouchier, Science **305**, 371 (2004).

[14] J. B. Plotkin, J. Dushoff, and S. A. Levin, Proc. Natl. Acad. Sci. USA (PNAS) **99**, 6263 (2002).

[15] A. C.-C. Shih, T.-C. Hsiao, M.-S. Ho, and W.-H. Li, Proc. Natl. Acad. Sci. USA (PNAS) **104**, 6283 (2007).

[16] R. Sanjuán, A. Moya, and S. F. Elena, Proc. Natl. Acad. Sci. USA (PNAS) **101**, 8396 (2004).

[17] S. Kryazhimskiy, J. Dushoff, G. A. Bazykin, and J. B. Plotkin, PLoS Genet **7**, e1001301 (2011), URL `http://dx.doi.org/10.1371%2Fjournal.pgen.1001301`.

[18] J. D. Bloom, L. I. Gong, and D. Baltimore, Science **328**, 1272 (2010).

[19] G. F. Rimmelzwaan, E. G. M. Berkhoff, N. J. Nieuwkoop, D. J. Smith, R. A. M. Fouchier, and A. D. M. E. Osterhaus, J Gen Virol **86**, 1801 (2005).

[20] M. E. J. Newman and R. Engelhardt, Proc. R. Soc. London B **256**, 1333 (1998).

[21] K. Koelle, S. Cobey, B. Grenfell, and M. Pascual, Science **314**, 1898 (2006).

[22] We identify the viral strain with its epitope sites by representing them consecutively in a unique connected sequence. The generalization to a four letters alphabet will of course modify the quantitative results reported here, but should not affect our qualitative conclusions.

[23] The choice of consecutive bits has been made by sake of simplicity, but any pair of sites could be chosen without loss of generality.

[24] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).