New methods for analyzing serological data with applications to influenza surveillance

Wilfred Ndifon*
Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544
October 1, 2009

Abstract

Two important challenges to the use of serological assays for influenza surveillance include the substantial amount of experimental effort involved, and the inherent noisiness of serological data. Here, informed by the observation that log-transformed serological data (obtained from the hemagglutination-inhibition assay) exist in an effectively one-dimensional space, computational methods are developed for accurately and efficiently recovering unmeasured serological data from a sample of measured data, and systematically minimizing noise found in the measured data. Careful application of these methods would enable the collection of better-quality serological data on a greater number of circulating influenza viruses than is currently possible, and improve the ability to identify potential epidemic/pandemic viruses before they become widespread. Although the focus here is on influenza surveillance, the described methods are more widely applicable.

Introduction

Serological data on circulating influenza viruses generally contain evolutionarily important information about functional (antigenic) variation in the B cell antigens of those

*Address (as of November 2009): Department of Immunology, the Weizmann Institute of Science, Rehovoth, Israel (ndifon@gmail.com).

viruses [1,2]. Because natural selection acts on antigenic variation, serological data can provide important insights into patterns, causes, and epidemiological consequences of influenza viral evolution [2-6]. Nevertheless, there are important challenges to the use of serological data. In particular, serological assays require considerable amounts of time and effort to perform, and this limits the number of viruses on which serological data can be routinely collected [7]. In addition, serological data are often contaminated by measurement noise (e.g., resulting from the serial dilution of sera), which may cause independently measured data for the same virus and serum sample to vary greatly. Furthermore, serological data depend on non-antigenic variables such as the red cell avidity and the antibody-inducing capacity of viruses [1,8], which can make it difficult to extract from the data accurate information about the antigenic variation of influenza viruses.

Smith et al. [2] recently made great progress towards addressing the above challenges, focusing on data obtained from the widely-used hemagglutination-inhibition assay. Those data are typically reported in the form of an m by n table (or matrix) H, with m the number of assayed viruses and n the number of sera used in the assay. Each H^{ij} in H (called the HI titer of virus i relative to serum j) is the reciprocal of the maximum dilution of serum j that can effectively neutralize virus i, i=1,…,m, j=1,…,n. Smith et al. used titers from multiple tables to construct low-dimensional embeddings ("antigenic" maps) of viruses and sera in which the Euclidean distance between virus i and serum j was approximated by $N^{ij}=\log_2(h_j/H^{ij})$, where h_j is the maximum titer found in the jth column of table H. (Use of base 2 in the log-transformation of h_j/H^{ij} reflects the fact that titers are typically measured using 2-fold dilutions of sera. Base 2 is also used here when log-transforming titers). The constructed antigenic maps allowed unavailable values of N^{ij} to be accurately predicted from distances measured on the maps,

increasing the amount of available data. Because the maps were constructed using data from multiple tables, data recovered from the maps are, in principle, less noisy than data obtained from individual tables.

Here, computational methods that extend the above antigenic-map approach into novel directions are introduced. More precisely, for a given m by n table H of titers, the methods enable accurate and efficient recovery of unmeasured values of H^{ij}, as well as the computation of confidence intervals (CIs) for measured values of H^{ij}, i=1,…,m, j=1,…,n. The methods can, in principle, also be used to directly recover suitably normalized titers, including N^{ij} (Ref. 2) and H^{ij}/H^{jj} (Ref. 7), as well as other measures of antigenic differences between viruses. In addition, the methods allow the minimization of both measurement noise and other types of non-antigenic variation found in titers, as well as the quantification of antigenic differences using such "noise-filtered" titers. The methods were partly motivated by remarkable recent work on the recoverability, from incomplete data samples, of data that exist in a low-dimensional space [9].

Results and discussion

Let H denote an m by n table of titers, with $m \geq n$. Because it is tedious to measure the mn titers found in H, it will be very helpful if only $s < mn$ titers can be measured and used to recover the unmeasured titers. The results of Candès and Recht [9] suggest that if the measured titers are selected randomly with uniform probability and $s \geq Crm^{1.2}\log(m)$, then the unmeasured titers can be recovered exactly if (i) $r \leq m^{1/5}$ and (ii) H has a low coherence (e.g., on the order of 1), where $C>0$ and r is the rank of H (see Methods for definitions). The unmeasured titers are recovered by finding an m by n table X that has minimal nuclear norm subject to the constraint that $X^{ij}=H^{ij}$, for all measured titers H^{ij} (Methods). To investigate the applicability of this method for recovering

titers, both the rank and the coherence of published tables of empirical titers [10] were computed (Methods). (Because missing titers can artificially lower the rank of a table, only complete tables were analyzed; see Supplementary Table 1). When the titers found in each table are log-transformed, on average ~99% of the variation in those titers is explained by the largest singular value of the considered table (Fig. 1a), suggesting that the analyzed tables have an effective rank of ~1. In addition, the tables have an average coherence of 1.36±0.12. These results suggest that unmeasured titers can be recovered exactly by the above method. Note, however, that because measured titers are likely to contain noise, exact recovery of unmeasured titers may not be possible. Also, recovery is not possible if entire rows/columns of H do not contain any measured titer. Furthermore, accurate (not necessarily exact) recovery is only possible if s is not smaller than the number of degrees of freedom of H, which equals $r(m+n-r)$ [9].

The above method for recovering unmeasured titers was applied to tables of empirical titers (Supplementary Table 1). The titers found in each table were log-transformed, and 90% of those titers were randomly selected and used to recover the unselected ("unmeasured") titers (Methods). This procedure was repeated 100 times for each table. [Note that it was necessary to limit this analysis to m by n tables for which $0.9mn \geq m^{1.2}\log(m)$]. Both the mean absolute difference (0.66±0.68) and the relative mean absolute difference (0.08±0.10) between the recovered and the unselected titers were small, suggesting that the recovery of unmeasured titers was accurate, albeit not exact. The correlation between the recovered and the unselected titers was high (R^2=0.77). A closer look at the results reveals that the mean absolute difference varies across the analyzed tables, ranging from 0.15 to 2.50. This may reflect the differing amounts of noise found in titers from different tables. To minimize such noise, a method was developed for computing noise-free estimates and CIs for titers (Methods). The method was tested using

20×10 rank-1 tables consisting of simulated, log-transformed titers. Each table was constructed as the product of two matrices (of dimensions 20×1 and 1×10, respectively), the entries of which were independently drawn from a truncated normal distribution with support (1.5, 4); this support was chosen to simulate the dynamic range of log-transformed, empirical titers, (3.32, 13.32). [Note that because empirical titers are measured on a geometric scale, it is reasonable to approximate their distribution by a lognormal distribution (e.g., Ref. 11)].

To simulate noise found in empirical titers, for each constructed table *M*, another table *M** was obtained by adding a perturbation (independently drawn from a normal distribution with mean 0 and standard deviation δ) to each entry of *M*. *M** was then used to compute estimates of and 95% CIs for the noise-free titers in *M* (Methods), for δ=0,0.1,0.2,…,1. Representative results obtained using one of the constructed tables are shown in Figure 1. The results show that the computed CIs have excellent coverage properties; they contain their corresponding noise-free titers in >99% of cases and they also have small relative widths (Fig. 1b & c). Remarkably, although estimates for the noise-free titers were computed using noisy titers the mean absolute difference between those estimates and the noise-free titers is much smaller than between the estimated and noisy titers, and between the noisy and noise-free titers (Fig. 1d). Indeed, the mean absolute difference between the estimated and noise-free titers grows much more slowly with δ than does the mean absolute difference between the estimated and noisy titers (Fig. 1d). These results suggest that the developed method can, in principle, be used to systematically minimize noise found in tables of empirical titers.

[Insert Figure 1]

In addition to measurement noise, it is also important to minimize other types of non-antigenic variation found in titers. This is currently not possible due to limited understanding of

the fundamental nature of titers. To shed light on the nature of titers, a mechanistic model of hemagglutination inhibition was developed and used to derive the first explicit mathematical equation for the titer of virus *i* relative to serum *j*:

$$H^{ij} = A^j K^{ij} J^i, \qquad (1)$$

where A^j denotes the concentration of antibodies found in serum *j*, K^{ij} the average affinity of those antibodies for virus *i*, and J^i a dimensionless quantity that depends on such non-antigenic variables as the avidity of virus *i* for red cell, the concentration of virus *i*, etc (see Supplementary Text for additional details). A^j depends on non-antigenic variables, including the antibody-inducing capacity of the virus against which serum *j* was raised, the immune status of the organisms in which serum *j* was raised, etc [8]. The derived equation predicts that the normalized titer H^{jj}/H^{ij}, a commonly used measure of antigenic difference (e.g., Ref. 7), is approximately independent of A^j, but it depends on both J^i and J^j. In contrast, a measure of antigenic difference introduced by Archetti and Horsfall [12] – $[H^{ii}H^{ij}/(H^{ij}H^{ji})]^{1/2}$ – is predicted to be approximately independent of the non-antigenic variables A^j, J^i, and J^j, suggesting that it may be more accurate. This is consistent with previous empirical results [10]. Importantly, the derived equation predicts that by mean-centering each row and column of a table of log-transformed, normalized titers the dependence of those titers on non-antigenic variables would be minimized. A method for quantifying and visualizing antigenic differences between viruses using such mean-centered tables is described in Methods.

In summary, the computational methods presented in this paper suggest new possibilities for improving the use of HI titers (and serological data in general) for influenza surveillance. In particular, the method for computing estimates and CIs for noise-free titers would allow uncertainties associated with titers to be taken into account, for example, when selecting viruses

for use in influenza vaccines. Also, the method for minimizing non-antigenic variation found in titers may help to improve the estimation of antigenic differences between viruses. In addition, the method for recovering unmeasured titers may help to reduce the experimental effort required to collect titers; for example, only $20^{1.2}\log(20)\approx 109$ titers may need to be measured in order to accurately determine all 200 possible titers for 20 viruses relative to 10 sera. The additional experimental capacity made available by this method would increase the number of viruses, circulating in humans and other organisms, for which titers can be routinely collected, and thereby improve the likelihood that potential epidemic and pandemic viruses will be identified before they become widespread (e.g., Ref. 7). This is important in light of the ongoing pandemic spread of an influenza virus whose initial circulation in humans may have gone undetected for several months [13].

Note that the observation, reported here, that tables of empirical titers have an effective rank of ~1 – titers exist in a space that has only one effective dimension – suggests that the effective number of independent variables that determine titers is very small. In other words, while titers depend on many variables [see Equation (1)] some of which may be mutually independent, variation in titers may be dominated by one or more co-dependent variables. In so far as titers are determined by and contain information about viral phenotypes, the demonstrated low dimensionality of the space of titers and the consequent recoverability of unknown titers suggests that it may be possible to predict biophysically accessible viral phenotypes from information about extant viral phenotypes. Well-designed theoretical and experimental tests of this idea may yield important new insights, and also shed light on questions concerning the evolutionary accessibility of epidemic variants of influenza viruses (see, e.g., Ref. 14). In addition to influenza surveillance, the new methods presented in this paper can be used in the

surveillance of other pathogens and in the investigation of basic questions concerning pathogen evolution and dynamics. As should be the case for existing methods (e.g., recently developed methods for elucidating the antigenic structure [2,15] and the adaptation potential [16] of pathogen populations), continued empirical/experimental assessment of the new methods is necessary to ensure that they will remain useful.

Acknowledgements: The author thanks Jonathan Dushoff for very helpful comments on a previous version of this manuscript; Leonid Kruglyak, Sergey Kryazhimskiy, Simon Levin, and Ned Wingreen for very helpful discussions; and the U.S. Centers for Disease Control and Prevention for making public the serological data used in this study.

Methods

Computing the effective rank and the coherence of a table of titers. Let $H = U * S * V^T$ be the singular value decomposition (SVD) of an *m* by *n* table *H* of log-transformed titers. The columns (called eigenvectors) of *U* (*V*) are orthonormal bases for the column (row) space of *H*, whereas the diagonal entries of *S* are the *n* singular values of *H*, $\lambda_i, i = 1,\ldots,n$, sorted in decreasing order of magnitude. ("*T*" denotes matrix transpose, and "*" denotes matrix multiplication). The fraction of the variation in titers that is explained by the *r* largest singular values of *H* is given by:

$$F_r = \frac{1}{\sum_{i=1}^{n}(\lambda_i)^2}\sum_{i=1}^{r}(\lambda_i)^2 . \qquad (2)$$

The rank of *H* is defined as the number of its non-zero singular values. Because some non-zero singular values may make negligible contributions to the variation in titers, the rank of *H* may be

larger than its effective rank, defined here as the smallest value of r for which $F_r \approx F_{r+1}$, $r=1,\ldots,n-1$ (the effective rank is set to n if this condition is not satisfied for any $r<n$).

Let P_U (P_V) be the orthogonal projection of H onto the first r columns of U (V). The "coherence" of H (more precisely, the maximum coherence of U and V) with respect to the standard bases in R^m and R^n is given by [9]:

$$\mu(H) = \max\left(\frac{m}{r}\max_{1\le i\le m}\left\|P_U e_i\right\|^2, \frac{n}{r}\max_{1\le i\le n}\left\|P_V e_i\right\|^2\right). \quad (3)$$

Recovering titers. Let H denote an m by n table of log-transformed titers of rank r, $m \ge n$, and let Ω denote a subset of $Crm^{1.2}\log(m)$ titers randomly selected from H with uniform probability, $C>0$. ($C=r=1$ is used in this study). The unselected (or "unmeasured") titers are recovered by finding an m by n matrix X that minimizes:

$$\mu\left\|X\right\|_* + \eta\sum_{H^{ij}\in\Omega}\left(H^{ij} - X^{ij}\right)^2, \quad (4)$$

where μ and η are Lagrange multipliers. $\left\|X\right\|_*$ denotes the nuclear norm of X, that is, the sum of the singular values of X. When r is known (it is not approximated by the effective rank of H), $\left\|X\right\|_*$ is replaced by the sum of the r largest singular values of X. Note that (4) is convex, so its optimal solution can usually be found efficiently [see Supplementary Text for details on the algorithm used to solve (4)].

Because titers recovered by the above method are theoretically exactly equal to their "noise-free" values when the selected titers are free of noise [9], discrepancies between the recovered and the noise-free titers are necessarily due to noise found in the selected titers. If there is no systematic bias in the way that noise found in the selected titers induces variation in

the recovered titers, then the recovered titers would be randomly distributed about the corresponding noise-free titers. The distribution of the recovered titers can therefore be used to compute CIs for the noise-free titers. This is the rationale for the following procedure for computing CIs for titers found in H: 1. Randomly select $m^{1.2}\log(m)$ titers from H. 2. Recover the unselected titers (see above). 3. Repeat steps 1 and 2 until each titer found in H is recovered at least N (=1000) times (on average it will take $nN/[n-m^{0.2}\log(m)]$ repetitions of steps 1 and 2 for this to happen). Let L_i be a list of the $k \geq N$ values recovered for the ith titer, which are sorted in increasing order. Then, the lower (upper) 95% CI for the ith titer is given by the $\lfloor .025k \rfloor^{th}$ ($\lceil .975k \rceil^{th}$) element of L_i, where $\lfloor x \rfloor$ ($\lceil x \rceil$) denotes the largest (smallest) integer smaller (larger) than x. The mean of the recovered values for each titer are used as a noise-free estimate for that titer.

Quantifying and visualizing antigenic differences between viruses. Let $H = U * S * V^T$ be the SVD of a table H of log-transformed titers. To quantify antigenic differences between viruses found in H, H is projected onto an r-dimensional subspace of its row space: $W^T = V_r^T * H^T$, where V_r denotes the first r columns of V. The antigenic difference between viruses i and j is defined as the Euclidean norm of the difference between the ith and jth rows of W. If $r \leq 3$, then antigenic differences can be visualized by plotting the rows of W. Antigenic differences computed using titers from different tables can be embedded in a common r-dimensional space by means of probabilistic multidimensional scaling (Supplementary Text). Note that the above SVD decomposition of H is only feasible when there are no missing titers in H (see Supplementary Text for a more generally applicable SVD approach).

References

1. Hirst GK (1941) The agglutination of red cells by allantoic fluid of chick embryos infected with influenza virus. Science 94: 22-23.
2. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, et al. (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305: 371-376.
3. Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, et al. (2008) The global circulation of seasonal influenza A (H3N2) viruses. Science 320: 340-346.
4. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, et al. (2009) Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325: 197-201.
5. Hancock K, Veguilla V, Lu X, Zhong W, Ebonèe N, et al (2009) Cross-reactive antibody responses to the 2009 pandemic H1N1 influenza virus. N Engl J Med. doi: 10.1056/nejmoa0906453.
6. Ndifon W, Wingreen NS, Levin SA (2009) Differential neutralization efficiency of hemagglutinin epitopes, antibody interference, and the design of influenza vaccines. Proc Natl Acad Sci USA 106: 8701-8706.
7. Layne SR (2006) Human influenza surveillance: the demand to expand. Emerg Infect Dis 12: 562-568.
8. Salk JE (1951) A critique of serologic methods for the study of influenza viruses. Arch Virol 4: 476-484.
9. Candès E, Recht B (2009) Exact matrix completion via convex optimization. Found Comp Math, doi:10.1007/s10208-009-9045-5.

10. Ndifon W, Dushoff J, Levin SA (2009) On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness. Vaccine 27: 2447-2452.

11. Reed GF, Lynn F, Meade BD (2002) Use of coefficient of variation in assessing variability of quantitative assays. Clin Diagnost Lab Immunol 9: 1235-1239.

12. Archetti I, Horsfall FL (1950) Persistent antigenic variation of influenza A viruses after incomplete neutralization *in ovo* with heterologous immune serum. J Exp Med 92: 441-461.

13. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, et al (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459: 1122-1125.

14. Recker M, Pybus OG, Nee S, Gupta S (2007) The generation of influenza outbreaks by a network of host immune responses against a limited set of antigenic types. Proc Natl Acad Sci USA 104: 7711-7716.

15. Plotkin JB, Dushoff J, Levin SA (2002) Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc Natl Acad Sci USA 99: 6263-6268.

16. Ndifon W, Plotkin JB, Dushoff J (2009) On the accessibility of adaptive phenotypes of a bacterial metabolic network. PLoS Comp Biol 5: e1000472.

Figure legend

Figure 1. Evidence for unidimensionality and high recoverability of titers. The dimensionality of each of 23 tables of empirical titers (Supplementary Table 1) was investigated by determining the fraction of the variation in (log-transformed) titers found in each table (denoted F_r) that is explained by the r largest singular values of that table (Methods), for r=1,…,5. To investigate the recoverability of titers, noise (independently drawn from a normal distribution with mean 0 and standard deviation δ) was added to each entry of a table consisting of simulated titers. The noisy titers were then used to compute estimates of and 95% CIs for the corresponding noise-free titers (Methods). (a) F_r (averaged over all 23 empirical tables) is plotted against r. (b) The fraction of noise-free titers that occurred within their computed CIs is plotted against δ. (c) The mean ratio of the width of the CI for a particular noise-free titer to the absolute value of that titer is plotted against δ. (d) The mean absolute difference between the estimated and noise-free titers, between the estimated and noisy titers, and between the noisy and noise-free titers, are plotted against δ. Bars denote standard deviations.

Figure 1