DOI:10.1111/j.1750-2659.2010.00192.x
www.influenzajournal.com

Original Article

New methods for analyzing serological data with applications to influenza surveillance

Wilfred Ndifon[a,b]

[a]Department of Immunology, The Weizmann Institute of Science, Rehovot, Israel. [b]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA.

Correspondence: Wilfred Ndifon, Department of Immunology, the Weizmann Institute of Science, P.O. Box 26, Rehovot, 76100 Israel. E-mail: ndifon@gmail.com

Accepted 9 November 2010. Published online 6 January 2011.

Background 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.

Results I show that log-transformed serological data exist in an effectively one-dimensional space. I use this result, together with new mechanistic insights into serological assays, to develop computational methods for accurately and efficiently recovering unmeasured serological data from a sample of measured data, for systematically minimizing noise and other types of non-antigenic variation found in the data, and for quantifying and visualizing antigenic variation. The methods can also be applied to data with effective dimensionality greater than one, under certain conditions.

Conclusion Careful application of the methods developed here 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 and pandemic viruses before they become widespread. Although the focus here is on influenza surveillance, the described methods are more widely applicable.

Keywords Data recovery, evolution, matrix completion, serological assays, surveillance.

Please cite this paper as: Ndifon W. (2011) New methods for analyzing serological data with applications to influenza surveillance. Influenza and Other Respiratory Viruses 5(3), 206–212.

Introduction

Serological data on circulating influenza viruses generally contain evolutionarily important information about functional (antigenic) variation in the B-cell antigens of those 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 potential epidemic and pandemic viruses on which serological data can be routinely collected.[7] In addition, serological data are often contaminated by experimental 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, the 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 obtain accurate information about the antigenic variation of influenza viruses.

Smith *et al.*[2] recently made great progress toward addressing the above challenges, focusing on data obtained from the widely used hemagglutination-inhibition assay. Those data are typically reported in a tabular format, with rows corresponding to viruses and columns to sera. The entry found in the i'th row and j'th column of the table (called the titer of virus i relative to serum j) represents the reciprocal of the maximum dilution of serum j that can effectively neutralize virus i. Smith *et al.* used titers from multiple tables to construct an 'antigenic' map of viruses and sera in which the Euclidean distance between virus i and serum j was approximated by the negative log (base 2) of the titer of virus i relative to serum j normalized by the maximum titer obtained for that serum. (Use of base 2 in the log-transformation of titers reflects the fact that titers are typically measured using twofold dilutions of sera. Base 2 is also used here when log-transforming titers). The constructed antigenic map allowed unavailable normalized titers to be accurately predicted from distances measured on the map, increasing the amount of available data. Because the map was constructed using data from multiple tables, data recovered from it are, in principle, less noisy than data obtained from individual tables.

Here, I introduce computational methods that extend the above antigenic-map approach into novel directions. First, I show that on average, log-transformed titers obtained in the same experiment (i.e. reported in the same table) have only one effective dimension. Then, I introduce a computationally efficient method that exploits this low dimensionality to accurately recover a complete table of titers using only a subset of those titers; various quantities derived from titers (e.g. the normalized titers of Ref. 2) can thus be recovered. I also show how 'noise-free' estimates and 95% confidence intervals (CIs) can be computed for titers and, hence, for related quantities. In addition, I derive a mathematical relation that sheds light on the mechanistic basis of titers and allows non-antigenic variation to be filtered out from these titers [Parts 1 and 2 of Supporting Information (SI)]. I introduce a method for quantifying and visualizing antigenic variation using such filtered titers, and I illustrate this method using titers for novel H1N1 viruses. The above methods were partly motivated by remarkable recent work on the recoverability of data that exist in a low-dimensional space.[9]

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 $m \times n$ table H of log-transformed titers. ('T' denotes matrix transpose and '*' denotes matrix multiplication). 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, λ_i, $i = 1, \ldots, n$, sorted in decreasing order of magnitude. The square of the i'th singular value (i.e. the i'th eigenvalue of $H^T H$) quantifies the variation explained by the i'th eigenvector;[10] therefore, the fraction of the variation explained by the first r eigenvectors is given by:

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

The rank of H is defined as the number of its non-zero singular values. Because some eigenvectors associated with non-zero singular values may explain negligible amounts of the variation in titers (e.g. because they contain noise), the rank of H may be greater than its effective rank, defined here as the smallest value of r for which $F_r = F_{r+1}$, $r = 1,\ldots,n-1$ (the effective rank is set to n if this condition is not satisfied for any $r < n$).

Alternatively, assuming that the noise found in log-transformed titers obtained in the same experiment is normally distributed, the eigenvectors containing noise would be associated with eigenvalues that are significantly smaller than the other eigenvalues.[11] In this case, the effective rank can be estimated by finding the smallest eigenvalue that is significantly greater than all the eigenvalues smaller than it, using Fisher's variance ratio test. This is equivalent to finding the maximum value of r for which[10,11]:

$$\frac{F_r - F_{r-1}}{1 - F_r}(n - r) > f_{1,n-r}(1 - \alpha), \quad (2)$$

where $f_{1,n-r}$ denotes the inverse cumulative function of the F-distribution with 1 and n-r degrees of freedom, and α is the level of statistical significance of the test. Here, I used $\alpha = 0.05/(n - r)$, where the factor (n-r) represents a Holm-Bonferroni correction for multiple comparisons.[12] Both of the above methods for determining the effective rank gave similar results.

Let P_U (P_V) be the orthogonal projection of H onto the first r eigenvectors found in 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]:

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

The lower the coherence, the greater the probability that measured titers, selected uniformly at random from H, will contain enough information about the unmeasured titers to enable their recovery by the method described elsewhere.

Recovering titers

Let H denote an $m \times n$ table of log-transformed titers of rank r, $m \geq n$, and let Ω denote a subset of $\sim rm^{1.2}\log(m)$ titers randomly selected from H with uniform probability. ($r = 1$ is used in this study). The unselected (or 'unmeasured') titers are recovered by finding an $m \times n$ matrix X that minimizes[13]:

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

where μ and η are Lagrange multipliers. $\|X\|_*$ 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), $\|X\|_*$ is replaced by the sum of the r largest singular values of X. Note that (Eqn. 4) was solved using the fixed point continuation algorithm[13]; for the experiments reported here, it took only a few seconds, on average, to recover titers found in each analyzed table using a Pentium IV machine running Windows XP.

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 $m^{1\cdot2}\log(m)$ 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 $nN/[n\text{-}m^{0\cdot2}\log(m)]$ 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 is used as a noise-free estimate for that titer. Note that if H is incomplete, then 95% CIs can still be computed for titers found in any $m_1 \times n_1$ complete sub-table of H, provided $m_1^{1\cdot2}\log(m_1) < m_1 n_1$.

In practice, when designing an experiment to measure titers for m viruses relative to n sera, the experimentalist can measure only a random subset of at least $m^{1\cdot2}\log(m)$ titers and subsequently recover the unmeasured titers as well as missing titers. In addition, noise-free estimates and 95% CIs for all titers (respectively subsets of titers) found in a complete (respectively incomplete) table can be computed and used in place of the original titers, which may be unreliable.

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 eigenvectors found in V. The antigenic difference between viruses i and j is defined as the Euclidean norm of the difference between the i'th and j'th 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 (Part 2 of SI). Note that the above SVD decomposition of H is only feasible when no titers are missing from H. In practice, missing titers should be recovered before H is subjected to SVD (see Part 2 of SI for an SVD approach that is applicable to tables with missing titers).

Results and discussion

Let H denote an $m \times n$ table of titers, with $m \geq n$. Because it is tedious to measure all mn titers found in H, it will be very helpful if fewer titers can be measured and used to recover the unmeasured titers. Recent theoretical work[9] suggests that if both the rank r and the coherence of H (see Methods for definitions) are not much greater than one, then unmeasured titers can be recovered exactly if only $\sim rm^{1\cdot2}\log(m)$ titers (selected randomly with uniform probability) are measured. I investigated the applicability of this method for recovering unmeasured titers by computing both the rank and the coherence of published tables of empirical titers[14] for influenza A and B viruses (Methods). (Because missing titers can artificially lower the rank of a table, only complete tables were analyzed; see Table S1). The analyzed tables, which typically consist of viruses sampled non-randomly from multiple influenza seasons, should contain more variation on average than tables consisting of viruses of the same subtype sampled randomly from only one season. Nevertheless, when the titers found in each table are log-transformed, on average ~99% of the variation in those titers is explained by only one eigenvector, which is associated with the largest singular value of the table (Figure 1A). This suggests that the analyzed tables have an average effective rank of ~1. I obtained a similar result by using a statistical F-test for the effective dimensionality of each table (Methods). In addition, the tables have an average coherence of 1·36 ± 0·12. Note that deviations from an effective rank of 1 may be caused by noise and other factors, including the normalization of titers (Part 3 of SI). Also, based on well-known results on the ranks of partitioned matrices,[15] when multiple tables of titers are combined the effective rank of the resulting composite table can be >1, especially if the rank of any one of the original tables is also >1 (Part 3 of SI). The above factors may partly explain the higher dimensionality of normalized titers reported previously.[2]

The fact that both the effective rank and the coherence of the analyzed tables are not much greater than one suggests unmeasured titers can be recovered exactly by the above-mentioned 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 for rows/columns of H that do not contain any measured titer. Furthermore, accurate (not necessarily exact) recovery is only possible if the number of measured titers is not smaller than the number of degrees of freedom of H, which is given by $r\,(m + n - r)$,[9] a condition that is satisfied by all the tables analyzed earlier. In practice, each row/column must contain more than r titers in order for titers missing from that row/column to be accurately recovered. It is important to keep in mind that exact recovery of unmeasured titers is still theoretically possible if the effective rank of H is >1, provided conditions (i) and (ii) above are satisfied. In addition to unmeasured titers, the above conditions are also applicable to the recovery of 'missing' titers – titers that are presumed to be either too

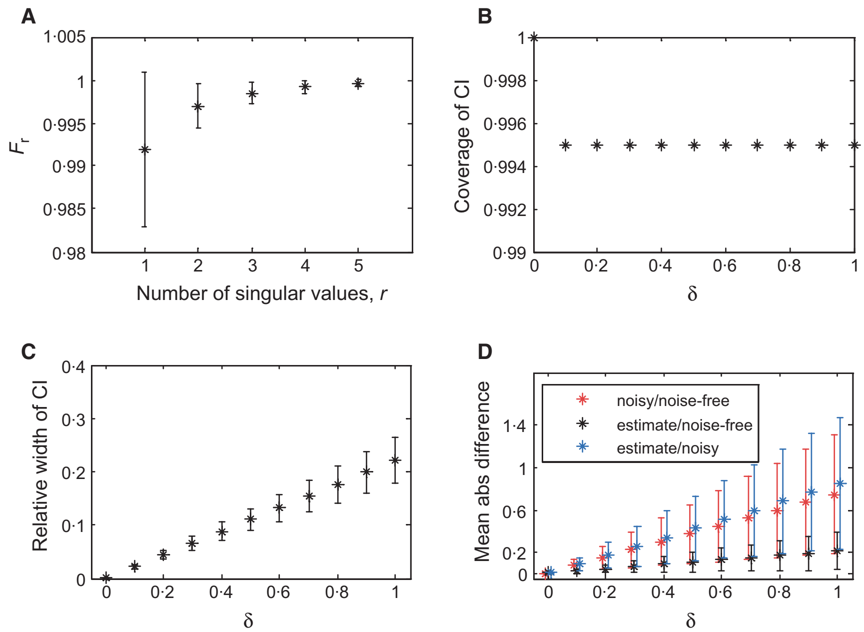

Figure 1. Dimensionality and recoverability of noise-free titers. (A) The dimensionality of each of 23 tables of empirical titers (Table S1) was investigated by log-transforming the titers found in each table and subsequently determining the fraction of the variation in titers (denoted F_r) that is explained by the eigenvectors associated with the r largest singular values of each table (Methods). F_r (averaged over all 23 tables) is plotted against r, for r = 1,…,5. (B) A table consisting of simulated titers was contaminated by noise (independently drawn from a normal distribution with mean 0 and standard deviation δ), and the resulting noisy titers were used to compute estimates of and 95% CIs for the uncontaminated (or noise-free) titers (Methods). The fraction of noise-free titers that occurred within their corresponding CIs is plotted against δ. (C) The mean ratio of the width of the 95% 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, is plotted against δ. Bars denote standard deviations.

high or too low to be measured experimentally. (Note that some missing titers may actually be within the limits of experimental resolution, but they could not be measured because of noise).

I applied the above method for recovering unmeasured titers to tables of empirical titers for influenza A and B viruses (Table S1). 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 $m \times n$ tables for which $0{\cdot}9mn \geq m^{1{\cdot}2}\log(m)$.] Both the mean absolute difference (0·66 ± 0·68) and the relative mean absolute difference (0·08 ± 0·10) between the recovered and the unmeasured titers were small, suggesting that the recovery of unmeasured titers was accurate, albeit not exact. A closer look at the above 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 noise found in empirical titers, I developed a novel method for computing noise-free estimates and 95% 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 log-normal distribution.[16]) 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. [Note that when δ = 1 there is a 32% chance that, on a geometric scale, a titer will be either multiplied or divided by at least 2 (the dilution factor often used in serological assays), representing a high amount of noise.] Representative results obtained using one of the constructed tables are shown in Figure 1.

The results show that the computed 95% CIs have excellent coverage properties; they contain their corresponding noise-free titers in >99% of cases and they also have small relative widths (Figure 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 (Figure 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 (Figure 1D). These results suggest that the developed method can, in principle, be used to systematically minimize noise found in tables of empirical titers. As an illustration, I computed noise-free estimates and 95% CIs for the empirical titers analyzed earlier. The average absolute difference between each titer and its noise-free estimate is small ($0{\cdot}66 \pm 0{\cdot}67$). However, 60% of the titers do not occur in their associated 95% CI, suggesting that those titers may be unreliable. This is consistent with the high variation in independently measured normalized titers for the same virus and antiserum that is observed in many cases (Table S1).

In addition to noise, it is also important to minimize other types of non-antigenic variation found in titers. This is currently not possible because of limited understanding of the fundamental nature of titers. To shed light on the nature of titers, I developed and used a mechanistic model of hemagglutination inhibition to derive the first explicit mathematical equation for the titer of virus i relative to serum j [Eqn. (S23) of SI]:

$$H^{ij} = A^{j} K^{ij} J^{i}, \tag{5}$$

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 Part 1 of SI for additional details). A^{j} depends on non-antigenic variables, including the antibody-inducing capacity of the virus against which serum j was raised and the immune status of the organisms in which that serum was raised.[8] The derived equation predicts that the normalized titer $H^{\,jj}/H^{\,ij}$, a commonly used measure of antigenic difference,[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[17] – $[H^{\,ii}H^{\,jj}/(H^{\,ij}H^{\,ji})]^{1/2}$ – 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.[14]

Importantly, the derived equation predicts that by mean-centering each row (resp. each row and column) of a table of log-transformed, un-normalized (resp. normalized) titers the dependence of those titers on non-antigenic variables would be minimized. I developed a method for quantifying and visualizing antigenic differences between viruses using such mean-centered tables (Methods). As an illustration, I used the method to construct a one-dimensional map showing the serological relationships among 123 H1N1 viruses, 118 of which are novel H1N1 viruses (Figure 2). The map shows that the serological responses of novel H1N1 viruses are on average much more similar to each other than to the corresponding responses of earlier H1N1 viruses, particularly a virus isolated in 1930 and another isolated in 1976 (Figure 2). In addition, the map suggests the existence of a well-defined direction of antigenic evolution of the viruses. Because 'serological' maps, such as the one described here, have only one dimension, they may facilitate computational prediction of short-term changes in the serological attributes of epidemic viruses.

In summary, the computational methods presented in this paper suggest new possibilities for improving the use of hemagglutination-inhibition titers (and serological data in general) for influenza surveillance. In particular, the

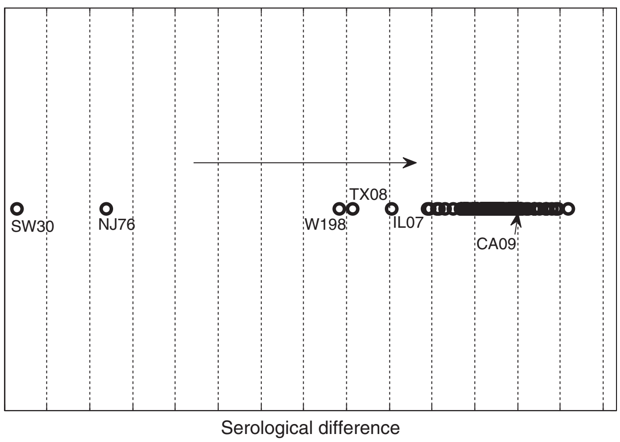

Figure 2. One-dimensional serological map of 123 H1N1 viruses. Four tables of titers (Supplementary Data S1) that were recently collected by the US. Centers for Disease Control and Prevention were analyzed. The tables had dimensions 26 by 10, 36 by 10, 40 by 12, and 41 by 11, respectively. Between 10% and 26% of titers were missing from each table. Measured titers found in each table were log-transformed and used to recover the missing titers, except for titers missing from rows/columns in which fewer than two titers were measured. (In particular, titers for the Brisbane/59/07 virus were not recovered). Each completed table was row-centered and used to quantify differences between the serological responses of pairs of viruses (Methods). These serological (or 'antigenic') differences were subsequently mapped onto a one-dimensional space by means of probabilistic multidimensional scaling (Part 2 of SI). The viral coordinates thus obtained are shown in the plot. The width of each displayed rectangle represents a unit of serological difference. The novel H1N1 vaccine virus (CA09) is labeled, as are five other viruses that were isolated before 2009. The arrow indicates the apparent direction of temporal changes in the serological properties of the analyzed viruses. Virus labels: CA09 California 07/2009, IL07 Illinois/9(33304)/2007, NJ76 New Jersey/8/1976, SW30 Swine/Iowa/1930, TX08 Texas/14/2008, WI98 Wisconsin/10/1998.

method for computing estimates and 95% 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 $20^{1.2}\log(20) = 109$ titers may need to be measured in order to accurately determine all 200 possible titers for 20 viruses relative to 10 sera. This method can be applied to other types of serological data, especially when the number of viruses under consideration is greater than or equal to the fifth power of the dimensionality of the data. The additional experimental capacity made available by the method would increase the number of viruses, circulating in humans and other organisms, for which serological data can be routinely collected and thereby improve the likelihood that potential epidemic and pandemic viruses will be identified before they become widespread.[7] This is important in light of the recent pandemic spread of an influenza virus whose initial circulation in humans may have gone undetected for several months.[18]

Note that the observation, reported here, that tables of empirical titers have an average effective rank of $\sim$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 (5) and Part 1 of SI] 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 high recoverability of unknown titers suggests that it may be possible to predict biophysically accessible viral phenotypes from limited 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.[19] 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[2,20] and the adaptation potential[21] of pathogen populations), continued empirical/experimental assessment of the new methods is necessary to ensure that they will remain useful.

Acknowledgements

I am grateful to Jonathan Dushoff and the anonymous reviewers for very helpful comments; Leonid Kruglyak, Sergey Kryazhimskiy, Simon Levin, and Ned Wingreen for very helpful discussions; and the US. Centers for Disease Control and Prevention for making public the serological data used here. I am also grateful to Ruth Arnon and Nir Friedman for their hospitality. This study was funded in part by a graduate fellowship from Princeton University, a DARPA grant (HR0011-05-1-0057), and a postdoctoral fellowship from the Weizmann Institute of Science.

References

1 Hirst GK. The agglutination of red cells by allantoic fluid of chick embryos infected with influenza virus. Science 1941; 94:22–23.

2 Smith DJ, Lapedes AS, de Jong JC *et al.* Mapping the antigenic and genetic evolution of influenza virus. Science 2004; 305:371–376.

3 Russell CA, Jones TC, Barr IG *et al.* The global circulation of seasonal influenza A (H3N2) viruses. Science 2008; 320:340–346.

4 Garten RJ, Davis CT, Russell CA *et al.* Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans. Science 2009; 325:197–201.

5 Hancock K, Veguilla V, Lu X *et al.* Cross-reactive antibody responses to the 2009 pandemic H1N1 influenza virus. N Engl J Med 2009; doi: 10.1056/nejmoa0906453.

6 Ndifon W, Wingreen NS, Levin SA. Differential neutralization efficiency of hemagglutinin epitopes, antibody interference, and the design of influenza vaccines. Proc Natl Acad Sci USA 2009; 106:8701–8706.

7 Layne SR. Human influenza surveillance: the demand to expand. Emerg Infect Dis 2006; 12:562–568.

8 Salk JE. A critique of serologic methods for the study of influenza viruses. Arch Virol 1951; 4:476–484.

9 Candès E, Recht B. Exact matrix completion via convex optimization. Found Comp Math 2009; doi:10.1007/s10208-009-9045-5.

10 Malinowski ER. Statistical *F*-tests for abstract factor analysis and target testing. J Chemom 1988; 3:49–60.

11 Carey RN, Wold S, Westgard JO. Principal component analysis: an alternative to ''referee'' methods in method comparison studies. Anal Chem 1975; 47:1824–1829.

12 Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979; 6:65–70.

13 Ma S, Goldfard D, Chen L. Fixed point and Bregman iterative methods for matrix rank minimization. Math Prog 2009; doi:10.1007/s10107-009-0306-5.

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

15 Marsaglia G, Styan GPH. Equalities and inequalities for ranks of matrices. Lin Multilin Algebra 1974; 2:269–292.

16 Reed GF, Lynn F, Meade BD. Use of coefficient of variation in assessing variability of quantitative assays. Clin Diagn Lab Immunol 2002; 9:1235–1239.

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

18 Smith GJD, Vijaykrishna D, Bahl J *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 2009; 459:1122–1125.

19 Recker M, Pybus OG, Nee S, Gupta S. The generation of influenza outbreaks by a network of host immune responses against a limited set of antigenic types. Proc Natl Acad Sci USA 2007; 104:7711–7716.
20 Plotkin JB, Dushoff J, Levin SA. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc Natl Acad Sci USA 2002; 99:6263–6268.
21 Ndifon W, Plotkin JB, Dushoff J. On the accessibility of adaptive phenotypes of a bacterial metabolic network. PLoS Comp Biol 2009; 5:e1000472.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. List of tables of empirical titers for influenza A and B viruses.

Table S1. Supporting text and list of tables of empirical titers.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.