***Science* Supplemental Online Material**

Mapping the Antigenic and Genetic Evolution of Influenza Virus

Derek J. Smith, Alan S. Lapedes, Jan C. de Jong, Theo M. Bestebroer,
Guus F. Rimmelzwaan, Albert D. M. E. Osterhaus, Ron A. M. Fouchier

published online 24 June 2004
doi: 10.1126/science.1097211

MATERIALS AND METHODS

Antigenic map construction

Influenza antigenic data. Within the framework of the WHO global influenza surveillance network, national influenza centers collect epidemic influenza virus isolates for antigenic characterization using the hemagglutination inhibition (HI) assay. The HI assay is based on the ability of the influenza virus hemagglutinin protein to cause agglutination of red blood cells (RBCs) and of specific antisera to block this reaction. These antisera are generally raised in ferrets by infection with prototypic strains, and a fixed quantity of each epidemic virus is tested against a panel of such reference antisera. Two-fold serial dilutions of the antisera are used to determine the highest dilution able to block the agglutination of RBCs. This dilution is called the HI titer, or HI value. The titers can range from <10 (lower dilutions are not used because of potential nonspecific inhibition, to limit the number of titrations, and to limit the amount of antiserum used in each test) to >10,240 (higher titrations are generally not used to limit the number of titrations). The higher the HI value the greater the antigenic relationship between the epidemic virus and antiserum.

Error function. We construct an antigenic map by minimizing the error function, $E=\sum_{ij} e(D_{ij},d_{ij})$, as a function of the antigen and antisera coordinates in the map, thereby determining the position of the antigens and antisera in the map. D_{ij} is the target distance between antigen i and antiserum j, derived from the HI measurement, H_{ij}, using the relation determined in (*S1*), $D_{ij} = b_j - log_2(H_{ij})$, with b_j set to the log_2 of the maximum measurement for antiserum j. d_{ij} is the Euclidean distance between the coordinates of antigen i and antiserum j in the map. $e(D_{ij},d_{ij})=(D_{ij}-d_{ij})^2$ when the HI measurement is a numeric value, and $e(D_{ij},d_{ij})=(D_{ij}-1-d_{ij})^2 g(D_{ij}-1-d_{ij})$ when the HI measurement is a thresholded value (typically <10). $g(x)=1/(1+e^{-10x})$ is a squashing function which allows $e(D_{ij},d_{ij})$ to contribute to E only if $d_{ij} < D_{ij}-1$. We minimize the error function using multiple random restarts of the conjugant gradient optimization method (*S2*).

Dimensionality. When constructing an antigenic map, one must not only determine the coordinates of the antigens and antisera, but also the dimensionality of the map. To determine the dimensionality, one typically constructs antigenic maps in different numbers of dimensions and then uses a variety of techniques to determine the number of dimensions that most suit the data. A stringent technique is the prediction of HI values which were missing from the HI data used to construct the map, and then measuring these predicted values using the HI assay—this was the method used here. Prediction error decreased only slightly as the number of dimensions increased from 2 to 5, and the cluster structure of the 2, 3, 4, and 5 dimensional maps were identical, indicating that distortion due to compression from higher dimensions to two dimensions is acceptably small, and that a 2-dimensional antigenic map is a good characterization of this data set.

Each of the dimensions in the higher dimensional maps represents antigenic distance just as in the lower dimensional maps. Mathematically, space is not limited to the usual three dimensions of which we are accustomed to thinking. The mathematical properties of space, such as the Euclidean distance between points, can be calculated in a unique and standard way for any number of dimensions.

The biochemical process of binding is complex (involving among other things complementarity in shape and charge, hydrogen bonding, and van der Waals interactions), and the dimensionality of the space necessary to describe binding has been estimated to be in the range of five to ten dimensions (*S3*). Thus the evolution of a pathogen such as human influenza A(H3N2), which we have shown to be mostly two-dimensional, can be thought of as a two-dimensional manifold (subspace) in a higher dimensional space. This dimension reduction is likely due to the interplay of many factors such as the ecological, epidemiological, and evolutionary processes which likely govern influenza strain dynamics (*S4–S7*). We expect that higher dimensional antigenic maps will be required for some pathogens, potentially corresponding to the pathogen's phylodynamics (*S7*).

Comparison with previous work. Most previous approaches to analyzing antigenic data have used the techniques of numerical taxonomy. The limitations of numerical taxonomy when applied to binding assay data are discussed in the main text and by Lapedes and Farber (*S1*). In addition, numerical taxonomy methods do not produce a metric space in which distance directly corresponds to individual measurements in the original data—instead distance is an aggregate of multiple measurements; thus, individual HI values cannot be directly predicted.

Lapedes and Farber introduced a fundamentally different approach, using a variant of ordinal multidimensional scaling (MDS) (*S8*, *S9*) to reconstruct a "shape-space" including both antigens and antisera (*S1*). This ordinal approach used an error function that was a function of distances in Euclidean space, d_{ij}, (which is in turn a function of coordinates of points in the space) and the *rank* of individual HI measurements. Although ordinal MDS is a *nonparametric* method, it produces a metric space. In this space, antigens and antisera are both represented and thus the results can be tested by prediction of individual HI values. One advantage of the ordinal approach is that it is not necessary to specify a priori the parametric relationship between HI value and distance, indeed this relationship can be determined using ordinal MDS and it turns out that distance is linearly related to the logarithm of the HI value (*S1*).

We have used this linear relationship as the foundation of a new *parametric* method to generate antigenic maps, and thus the relative positions of antigens and antisera in the map can be determined with a variant of metric (instead of ordinal) MDS (*S10*, *S11*). Metric approaches construct an error function that is a function of the distances in Euclidean space and the *value* of individual HI measurements (instead of the *rank* of the HI measurements used in ordinal approaches).

When an HI measurement is beyond the sensitivity of the assay, it is reported as a thresholded value, typically "<10." To account for such relational information we modified the metric approach to be able to deal with both values and relations, as described in the section on the error function above.

Consensus. The value of the error function for the antigenic map of Fig. 1 is 3523.74. Consensus among the different minimizations was high; the relative positions of pre-TX77 strains was highly repeatable, as was the relative positions of post-TX77 strains; however, the relative position of the pre-TX77 region to the post-TX77 region had more than one stable state (Fig. S2). Distances between pre- and post-TX77 strains were excluded when calculating the correlation between antigenic distance and the number of aa substitutions due to this metastability.

Accuracy. Since antigen-antiserum distances in the map correspond to HI values, it was possible to predict HI values that were missing in the original data set, and subsequently measure those values using the HI assay, to determine the resolution of the map. When we blindly predicted 481 such HI measurements, the average magnitude of the prediction error for the two-dimensional map was 0.83 [SD 0.67, correlation between predicted and measured values 0.80 ($P << 0.01$)] units. In addition to blind prediction from a 2-dimensional map, we also blindly predicted from 3, 4, and 5 dimensional maps. Prediction error decreased only slightly as the number of dimensions increased, indicating (along with previous evidence that the cluster structure of the 2, 3, 4, and 5 dimensional maps were the same) that distortion due to compression from higher dimensions to 2 dimensions is acceptably small for this data set.

Rigidity. The rigidity of the coordinates of points in an antigenic map depends not only on the number of HI measurements between the antigens and antisera, but also on how well spread the antigens and antisera are in space. In direct analogy to geographic map making, the coordinates of points can be determined most accurately if they can be well triangulated with other points. For example, if three columns of antisera have similar HI values for each of the antigens they are measured against, then those antisera will likely end up in similar locations in the antigenic map, and although they will provide repeated-measure type information to determine the antigenic distance between the antigens and the three antisera, they will provide little information about the distances among the antigens. In contrast, three antisera that are more widely spread will be able to locate the antigens relative to each other, as well as to the antisera. Thus when selecting reference antisera, it is best to choose antisera which are distinct from each other. Once an antigenic map has been made, the accuracy (by prediction described above), rigidity of individual points (described below), and sensitivity to sample by bootstrapping (described below) can be derived and confidence areas delineated; if required, extra HI values can be measured to increase the resolution and tighten the confidence areas. Given antisera and antigens that are well spread in space, an antigenic map with accuracy similar to the one shown in Fig. 1 can be made from relatively small amounts of data. For example, most of the HI tables used in this study contain data predominantly from a single antigenic cluster, and contain antigens measured against typically 6 to 8 antisera, and almost all of these HI tables produce reasonable antigenic maps even when not combined into a single data set.

To determine the rigidity confidence area for each individual antigen and antiserum, all other antigens and antisera were held fixed, and the target antigen or antiserum was displaced to determine how the error function changed due to the displacement. The periphery of the shapes in Fig. 1 indicates an increase of 0.5 in the error function.

Shading illustrates the rate of error increase for each antigen, from black (no error) to the base color of the antigenic cluster (0.5 error) at the periphery.

Sensitivity to the data set sample. We tested the sensitivity of the antigenic map to the particular antigens, antisera, and individual HI measurements by bootstrap resampling of each of these components. We used the standard bootstrap technique of sampling the existing data set with replacement and generating new data sets of the same size as the original data set typically resulting in resamples with multiple entries of some components and missing entries of other components. Other than the same trans-TX77 metastability discussed above, the maps generated from these new data sets were generally robust to such resampling. We also bootstrap resampled the antigens to generate data sets with the same number of antigens per season (separate resamples with 3 and 10 strains per season). Again, apart from the trans-TX77 metastability, the maps were generally robust to the resampling.

One of the major topographical features of the map in Fig. 1 is the change in direction of the antigenic evolution between the 1980s and the 1990s. The timing of this change is somewhat concomitant with the appearance of the BE89 cluster genetic dead-end. To test if this change was an artifact introduced by the BE89 cluster, we remade the map omitting the BE89 antigens and antisera. The resulting map was almost identical to the original, indicating that the change is not an artifact of the BE89 genetic dead-end.

Combining HI data. When combining HI data, either from tables made at different times in the same laboratory, or tables made in different laboratories, the map made from the combined data might have higher or lower resolution than maps made from the individual tables. The accuracy of the individual and combined maps, and individual antigens and antisera in the maps, can be determined by prediction (either by measuring previously unmeasured values, or by leaving out already measured values), calculating rigidity and bootstrap resampling confidence areas as described above. Unless there are systematic differences between the data sets, adding more data will increase the resolution of the map; if there are systematic differences, then they can be detected using these methods. These methods also provide a way to curate an antigenic database and maintain consistency: as new antigenic data are produced, reference antigens and/or antisera can be included, the positions of which are compared in the new and existing maps.

Variation in red blood cells (RBCs) and antisera. Variation in the reagents used in the HI assay, such as the animal source of red blood cells (RBCs), and animal used to raise the antisera, add potential complexity to HI data.

The data used in this study were collected in a single laboratory, all antisera were raised in ferrets, RBCs up to and including the 1991/92 season were from chickens and thereafter from turkeys. As part of the data validation process for this study, we remeasured, using turkey RBCs, 88 titers of strains from 1968 to 1989 that had previously been measured as part of routine surveillance using chicken RBCs. The average titer difference was 0.01 of a 2-fold dilution, and the average of the absolute values of the titer differences was 0.68 of a 2-fold dilution, indicating that the switch from chicken to turkey RBCs is unlikely to have a significant effect in this data set.

Had the switch from chicken to turkey RBCs had a larger effect, the methods described here would provide techniques to examine the differences quantitatively: a map made

with strains only measured using chicken RBCs could be compared to a map with measurements using turkey RBCs, and strains in common between the maps could be compared for differences in antigenic distance to other strains.

Similarly, maps could be made for the same antigens using antisera from different animals to analytically compare differences among such antisera. Hyperimmune sheep (or goat) antisera for example are less specific than ferret antisera; hence, an antigenic map made only with sheep antisera would be scaled differently from a map made only with ferret antisera, and the difference in scale could be determined by comparing the two maps. Furthermore, this difference in scale could then be used to make antigenic maps from HI tables containing measurements using both ferret and sheep antisera (the scale factor would be applied in the $e(D_{ij},d_{ij})$ part of the error function, which would take a third parameter to indicate the animal source of antisera). Alternatively, the scale factor could be a free parameter in the error function which would be then be determined as part of the error minimization process. If the difference between hyperimmune sheep or goat antisera were not linear, but varied as a function of the antigenic distance between antigens, then this could also be determined using the same methods, and the $e(D_{ij},d_{ij})$ term in the error function could be extended with a nonlinear function instead of a simple scale factor. In practice, it is likely that once the relation among ferret and hyperimmune sheep or goat antisera has been determined using these methods, it could then be used transparently in future analyses.

Genetic map construction

Our method for making genetic maps is similar to that for making antigenic maps, except that the target distances are the number of aa substitutions between the antigen aa sequences, and there are no antisera points. To avoid underestimating genetic distances due to sequential mutations at the same location, when a target distance was greater than 30 aa substitutions it was allowed to contribute to the error function only when the map distance was <30 in the same way that a thresholded HI measurement was only allowed to contribute to the error function when the map distance was within threshold. PCOORD (*S12*) is a related method for making genetic maps, but does not directly handle thresholded distances. The correlation between the number of aa substitutions and the corresponding distances between strains in the genetic map was 0.98 indicating that the 2D genetic map is a reasonable representation of the target aa distance matrix.

Clustering

Clusters were identified in the antigenic map by a *k*-means clustering algorithm (*S13*), using average weighting, and k=10. Two adjustments were then made. First, the three TX77-like strains were made into a separate cluster. TX77 is a well-known antigenic and genetic type (*S14*), and strains of this type were recorded in our laboratory, but apart from the three strains in the data set, the remaining strains, and HI data for these strains, are currently lost. Second, strains VI/7/87, OK/5/88, and VI/1/88 were assigned to the SI87 cluster—in multiple runs of the *k*-means clusterer these strains appeared in either the SI87 or BE89 clusters with very small error difference between placement in either cluster. The most parsimonious placement of these strains, based on genetic and temporal information is in the SI87 cluster.

When clusters were determined from the genetic map (instead of the antigenic map), and plotted season-by-season, the antigenic evolution was still punctuated and the genetic evolution still gradual.

"Cluster-difference" aa substitutions

An aa substitution X to Y at location L is considered a "cluster-difference" substitution between clusters A and B if all (or all but one) strains in cluster A have aa X at location L, and all (or all but one) strains in cluster B have aa Y at location L. Fig S1 shows an alignment of the aa sequences used in this study, grouped and color-coded based on antigenic cluster, and marked with the cluster-difference aa substitutions.

Evolutionary distance to A/Bilthoven/16190/68 antigen

One way to measure the distance of a strain to the root of the 1968 pandemic strains would be to measure the straight-line distance on the map from the strain of interest to a reference strain chosen to be representative of the 1968 pandemic. An obvious choice for such a reference strains would be the A/HongKong/1/68 (HK/1/68) vaccine strain. However, we choose a different strain, A/Bilthoven/16190/68 (BI/16190/68), because in both the ML tree (Fig. 2A) and a protein parsimony tree (data not shown) BI/16190/68 was closest to the avian outgroup, and also because with only 5 passages in tertiary monkey kidney cells and no passages in eggs, it likely has fewer laboratory adaptations than the HK/1/68 vaccine strain and thus is likely to be a better representative of the early pandemic strains. We also measure distance to BI/16190/68, not in a straight-line, but as an evolutionary distance through the centroids of clusters between the strain of interest and BI/16190/68. This path between cluster centroids is uniquely determined for influenza virus as the clusters are chronologically ordered (for other pathogens this route might not be clear). Note, this implies going via the BE89 cluster when passing from the BE92 cluster back to the SI87 cluster, as this is the order in which the clusters appeared, and thus the order which will influence selection at the phenotypic level and the most appropriate measure here of antigenic distance, rather than following the genetic lineage which would bypass the BE89 genetic dead-end. Similarly, the antigenic path from TX77 to EN72 goes through VI75. Taking this evolutionary path, rather than the straight-line distance, to the root antigen is a map analog of measuring genetic distance from a phylogenetic tree rather than as the number of mutations to the root strain—for short distances these measures are usually similar, but for longer distances the straight-line distance will be shorter than the evolutionary distance. The evolutionary distance to the BI/16190/68 strain in the antigenic and genetic maps was thus calculated as the sum of distances of a strain to the centroid of the previous antigenic cluster, from this centroid through the intermediate centroids to the EN72 centroid, and then to BI/16190/68 antigen in the HK68 cluster.

Biases in the data

There are three biases in the data used for this study. First, the majority of strains selected for sequencing as part of routine surveillance were those with unusual antigenic properties. The remaining strains were the first and last isolate from each season, a representative of each antigenically distinct cluster seen in the Netherlands that season (determined each year as part of routine surveillance by classical reading of HI data), and the vaccine strains. This sample is thus likely to contain a larger proportion of outliers than would be obtained from a random sample; thus, the cluster structure of a map made

from a random sample would likely be more clustered than the map of Fig. 1, and the observation of clustered antigenic evolution would be unlikely to change. The removal of outlier strains would be unlikely to alter the overall structure of the map as the map is robust to bootstrap resampling of the antigens and to the removal of whole clusters such as BE89. For a subset of seasons we generated an antigenic map for all strains collected at the Dutch National Influenza Center—this is closer to a random sample than the sequenced strains; the cluster structure of Fig. 1, and the punctuated antigenic evolution of Fig. 4A persisted. As indicated in the caption of Fig. 4, since sampling is biased towards outlier strains, the area of the circles in Figs. 4 A, B and C, does not reflect epidemic impact. The strains in each antigenic cluster were bootstrap resampled to test the robustness of distances between antigenic clusters in the genetic and antigenic maps shown in Table 1. Distances were not statistically significantly different from the distances in original sample, and SEs were sufficiently small to indicate that distances are robust to the surveillance sampling biases (Table S1).

Second, approximately one third of the strains in this study were isolated in a single country—The Netherlands. This sampling bias is likely to affect the season in which the first strains in a cluster are detected, as new variants are likely to circulate locally before their global emergence. A further effect is that some variants, which only circulate locally before dying out, might not appear in this data set. The addition or absence of such local data, as with the removal of outlier strains, would be unlikely to change the overall structure of the map as both the antigenic and genetic maps are robust to the presence or absence of even major clusters.

Third, mutations can be introduced by laboratory propagation of influenza virus strains in embryonated chicken eggs. However, since 89% of strains used in this study were propagated only in mammalian cell culture, for at most 5 passages, and usually only 1 or 2 passages, this bias is unlikely to have a significant effect on this study (*S15, S16*).

Accession numbers

In addition to the newly sequenced HA1 domains, the following previously published HA1 domain sequences were used in this study: AF008665 AF008697 AF008711 AF008725 AF008755 AF008769 AF008828 AF008867 AF008886 AF008888 AF008903 AF008905 AF092062 AF131997 AF180570 AF180602 AF180643 AF201874 AF368444 AF368446 D21173 D49961 ISDN38157 ISDNCDA001 ISDNENG72 ISDNHK71 ISDNTX77 ISDNVIC75 M16739 U08858 Z46405 Z46408 Z46413 Z46414.

Table S1. Distances between antigenic clusters in the genetic and antigenic maps. Data are related to the data in Table 1 of the main text except that distances and ratios are the average of 1,000 bootstrap resamples of the strains in each antigenic cluster, and the standard error (SE) of the averages are shown. Each resample was of the same number of strains per cluster as in the original sample. Distances and ratios differ slightly, but not statistically significantly, from those in Table 1 because these are averages of 1,000 bootstrap resamples, whereas the data in Table 1 were calculated from the original sample.

Cluster transition	Genetic distance (aa substitutions)		Antigenic distance (units)		Genetic/antigenic ratio	
	Av	SE	Av	SE	Av	SE
HK68-EN72	12.19	0.93	3.43	0.23	3.57	0.26
EN72-VI75	14.71	0.54	4.38	0.24	3.37	0.25
VI75-TX77	14.87	0.98	3.41	0.66	4.48	0.65
TX77-BA79	16.08	2.04	3.34	0.27	4.82	0.56
BA79-SI87	11.99	1.11	4.93	0.31	2.44	0.2
SI87-BE89	6.87	0.5	4.60	0.27	1.50	0.12
BE89-BE92	13.73	0.41	7.77	0.19	1.77	0.07
BE92-WU95	9.89	0.75	4.59	0.21	2.17	0.16
WU95-SY97	16.07	0.86	4.69	0.17	3.43	0.23
SY97-FU02	16.05	0.77	3.52	0.35	4.60	0.44
Average SE		0.89		0.29		0.29
Total	132.43	1.79	44.66	0.97		
Av	13.24	0.18	4.47	0.1	3.21	0.13

Figures appear at the endn of the file

Fig S1. Alignment of aa sequences used in this study, grouped and color-coded according to the antigenic clusters, and with the cluster-difference substitutions boxed. The nine strains interdigitated between the BE92 and WU95 clusters are underlined. Codons in which there were no aa substitutions are excluded. The image can be zoomed to show details.

Fig S2. The antigenic map of influenza A virus strains as shown in Fig. 2C, with the rigid-body rotation and translation of the pre-TX77 clusters shown by the arrow. The positions in the original map (Fig. 1) are shown in light colors, and darker colors indicate the positions after the rotation and translation.

Fig S3. The antigenic map of influenza A virus strains as shown in Fig. 1 except that virus strains are represented by colored circles and antisera by open squares. A line connecting a virus strain an antiserum indicates that the virus strain was used to raise the antiserum.

References and Notes

S1. A. Lapedes, R. Farber, *J Theor Biol* **212**, 57-69 (Sep 7, 2001).
S2. B. Flannery, S. Teukolsky, W. Vetterling, *Numerical Recipes in C* (Cambridge Univ. Press, Cambridge, 1988).
S3. A. S. Perelson, G. F. Oster, *J Theor Biol* **81**, 645-70 (1979).
S4. N. M. Ferguson, A. P. Galvani, R. M. Bush, *Nature* **422**, 428-33 (2003).
S5. J. R. Gog, B. T. Grenfell, *Proc Natl Acad Sci U S A* **99**, 17209-14 (2002).
S6. J. R. Gog, G. F. Rimmelzwaan, A. D. Osterhaus, B. T. Grenfell, *Proc Natl Acad Sci U S A* **100**, 11143-7 (2003).
S7. B. T. Grenfell *et al.*, *Science* **303**, 327-332 (2004).
S8. J. Kruskal, *Psychometrika* **29**, 1-27 (1964).
S9. J. Kruskal, *Psychometrika* **29**, 115-129 (1964).
S10. R. Shepherd, *Human Factors* **5**, 33-48 (1963).
S11. R. Shepherd, *J Math Psychol* **1**, 54-87 (1964).
S12. D. G. Higgins, *Comput Appl Biosci* **8**, 15-22 (1992).
S13. R. Duda, P. Hart, *Pattern Recognition and Scene Analysis* (Wiley Interscience, 1973).
S14. G. W. Both, M. J. Sleigh, N. J. Cox, A. P. Kendal, *J Virol* **48**, 52-60 (1983).
S15. J. C. de Jong *et al.*, *Bull World Health Organ* **66**, 47-55 (1988).
S16. G. C. Schild, J. S. Oxford, J. C. de Jong, R. G. Webster, *Nature* **303**, 706-9 (1983).