Geometric constraints dominate the antigenic evolution of influenza H3N2 hemagglutinin

Austin G. Meyer[1,2] and Claus O. Wilke[1*]

1 Department of Integrative Biology, Institute for Cellular and Molecular Biology, and Center for Computational Biology and Bioinformatics. The University of Texas at Austin, Austin, TX 78712, USA.
2 School of Medicine, Texas Tech University Health Sciences Center, Lubbock, TX 79430, USA.

∗ **Corresponding author: wilke@austin.utexas.edu**

Keywords: influenza, hemagglutinin, immune epitope, evolution

Abstract

We have carried out a comprehensive analysis of the determinants of human influenza A H3 hemagglutinin evolution, considering three distinct predictors of evolutionary variation at individual sites: solvent accessibility (as a proxy for protein fold stability and/or conservation), experimental epitope sites (as a proxy for host immune bias), and proximity to the receptor-binding region (as a proxy for protein function). We have found that these three predictors individually explain approximately 15% of the variation in site-wise dN/dS. However, the solvent accessibility and proximity predictors seem largely independent of each other, while the epitope sites are not. In combination, solvent accessibility and proximity explain 32% of the variation in dN/dS. Incorporating experimental epitope sites into the model adds only an additional 2 percentage points. We have also found that the historical H3 epitope sites, which date back to the 1980s and 1990s, show only weak overlap with the latest experimental epitope data, and we have defined a novel set of four epitope groups which are experimentally supported and cluster in 3D space. Finally, sites with $dN/dS > 1$, i.e., the sites most likely driving seasonal immune escape, are not correctly predicted by either historical or experimental epitope sites, but only by proximity to the receptor-binding region. In summary, proximity to the receptor-binding region, rather than host immune bias, seems to be the primary determinant of H3 immune-escape evolution.

Author summary

The influenza virus is one of the most rapidly evolving human viruses. Every year, it accumulates mutations that allow it to evade the immune response of people previously infected. Which sites in the virus' genome allow this immune escape is not entirely understood, but conventional wisdom states that specific "immune epitope sites" in the protein hemagglutinin are primarily responsible, and these sites are commonly targeted by vaccine development efforts. Here, we survey all available experimental data on immune epitopes in hemagglutinin, and we demonstrate that these immune epitope sites may not be that crucial for influenza evolution. Instead, we propose and find evidence for a simple geometrical model: sites that are closest to the location where the virus binds the human receptor (i.e, the furthest away from the viral surface) are the primary driver of immune escape.

Introduction

The influenza virus causes one of the most common infections in the human population. The success of influenza is largely driven by the virus's ability to rapidly adapt to its host and escape host immunity. The antibody response to influenza virus is determined by the surface proteins hemagglutinin (HA) and neuraminidase (NA). Among these two proteins, hemagglutinin, the viral protein responsible for receptor-binding and uptake, is a major driver of host immune escape by the virus. Previous work on hemagglutinin evolution has shown that the protein evolves episodically [1–3]. During most flu seasons, hemagglutinin experiences mostly neutral drift around the center of an antigenic sequence cluster; in those seasons, it can be neutralized by similar though not identical antibodies, and all of the strains lie near each other in antigenic space [4–7]. After several seasons in a cluster, the virus escapes its local sequence cluster to establish a new center in antigenic space [7–9].

There is a long tradition of research aimed at identifying the regions of hemagglutinin to which the human immune system responds, and by proxy, the sites that determine sequence-cluster transitions [4, 6, 10–21]. Initial attempts to identify and categorize immune-epitope sites of H3 hemagglutinin were primarily sequence-based and focused on substitutions that took place between 1968, the emergence of the Hong Kong H3N2 strain, and 1977 [10, 11]. Those early studies used the contemporaneously solved protein crystal structure, a very small set of mouse monoclonal antibodies, and largely depended on chemical intuition to identify antigenically relevant amino-acid changes in the mature protein. Many of the sites identified in those studies reappeared nearly two decades later, in 1999, as putative epitope sites with no additional citations linking them to immune data [4]. Those sites and their groupings are still considered the canonical immune epitope set today [3, 16, 22]. While the limitations of experimental techniques and of available sequence data in the early 1980's made it necessary to form hypotheses based on chemical intuition, these limitations have been overcome through recent advances in experimental techniques and wide-spread sequencing of viral genomes. Therefore, it is time to revisit the question of which sites in influenza hemagglutinin are epitope sites. In particular, since the original epitope set was identified via sequence analysis, we do not even know whether *bona-fide* immune-epitope sites actually exist, i.e., sites which represent a measurable bias in the host immune response. And even if such sites do exist, we do not know to what extent the bias in host immune response can be quantified with the available biochemical assays. Finally, assuming that immune-epitope sites exist and can be experimentally identified, it is possible that they do not experience an elevated selective pressure to change amino acids relative to other important sites in the protein.

Some recent studies have begun to address these questions indirectly, via evolutionary analysis. For example, over the last two decades, virtually every major study on positive selection in hemagglutinin has found some but never all of the historical epitope sites to be under positive selection [3, 16, 18, 19, 23]. Furthermore, each of these studies has found a set of sites that are under positive selection but do not belong to any historical epitope. Finally, because every study identifies slightly different sites, there seems to be no broad agreement on which sites are under positive selection [12, 16, 18, 19]. The sites found by disparate techniques are similar but they are never identical.

To dissect the determinants of hemagglutinin evolution and the evidence for epitopes, we linked several predictors, including relative solvent accessibility, the inverse distance from the receptor-binding region, and experimental immune epitope data, to site-wise evolutionary rates calculated from all of the human H3N2 sequence data for the last 22 seasons (1991–2014). We found that the inverse distance from the sialic acid-binding region explained the largest portion of evolutionary rate variation. Moreover, we analyzed all of the available H3 experimental epitope data, and we found that experimental data supports redefining and regrouping immune epitope sites for hemagglutinin. After controlling for biophysical constraints with relative solvent accessibility, function with distance to the receptor-binding region, and immune bias with experimental epitope data, the remaining explanatory power of historical categories was relatively low. Finally, by explicitly accounting for each constraint we found that we could predict nearly 35% of the evolutionary rate variation in hemagglutinin, nearly twice as much variation as could be explained by earlier models.

Results

Relationship between evolutionary rate and inverse distance to the receptor-binding site

Our overarching goal in this study was to dissect the factors that determine selective pressures at individual sites in influenza hemagglutinin H3. In particular, we wanted to identify specific biophysical or biochemical properties of the mature protein that determine whether a given site will evolve rapidly or not. As a measure of evolutionary variation and selective pressure, we used the metric dN/dS. dN/dS can measure both the amount of purifying selection acting on a site (when $dN/dS \ll 1$ at that site) and the amount of positive diversifying selection acting on a site (when $dN/dS \gtrsim 1$). For simplicity, we will refer to dN/dS as an *evolutionary rate*, even though technically it is a *relative* evolutionary rate or evolutionary-rate ratio. We built an alignment of 3854 full-length H3 sequences spanning 22 seasons, from 1991/92 to 2013/14. We subsequently calculated dN/dS at each site, using a fixed-effects likelihood (FEL) model as implemented in the software HyPhy [24].

Although a number of predictors of site-specific evolutionary rate variation have been developed previously [19, 20, 25–30] (see Ref. [31] for a recent review), none describe a functional constraint on the evolution of viral proteins. In addition, they generally all predict the amount of purifying selection expected at sites, and therefore they cannot identify sites under positive diversifying selection. Moreover, the short divergence time of viruses causes the systematic biophysical pressures that predict much of eukaryotic protein evolution to be much less dominant in viral evolution [28]. Thus, we set out to find a constraint on hemagglutinin evolution that was related to the protein's role in viral binding and fusion. A few earlier studies had shown that sites near the sialic acid-binding region of hemagglutinin tend to evolve more rapidly than the average for the protein [4, 20, 21]. Furthermore, when mapping evolutionary rates onto the hemagglutinin structure, we noticed that the density of rapidly evolving sites seemed to increase somewhat towards the receptor-binding region (Fig. 1A). Therefore, as the primary function of hemagglutinin is to bind to sialic acid and induce influenza uptake, we reasoned that distance from the receptor-binding region of HA might serve as a predictor of functionally driven HA evolution. We calculated distances from the sialic acid-binding region (defined as the distance from site 224 in HA), and correlated these distances with the evolutionary rates at all sites. We found that distance from the receptor-binding region was a strong predictor of evolutionary rate variation in hemagglutinin (Pearson correlation $r = 0.41$, $P < 10^{-15}$).

Next, we wanted to verify that this correlation was representative of hemagglutinin evolution and not just an artifact of the specific site chosen as the reference point in the distance calculations. It would be possible, for example, that distances to several spatially separated reference sites all resulted in similarly strong correlations. We addressed this question systematically by making, in turn, each individual site in HA the reference site, calculating distances from that site to all other sites, and correlating these distances with evolutionary rate. We then mapped these correlations onto the structure of hemagglutinin, coloring each site according to the strength of the correlation we obtained when we used that site as reference in the distance calculation (Fig. 1B). We obtained a clean, gradient-like pattern: The correlations were highest when we calculated distances relative to sites near the receptor-binding site (with the maximum correlation obtained for distances relative to site 224), and they continuously declined and then turned negative the further we moved the reference site away from the apical region of hemagglutinin (Fig. 1B). This result was in stark contrast to the pattern we had previously observed when mapping evolutionary rate directly (Fig. 1A). In that earlier case, while there was a perceptible preference of faster evolving sites to fall near the receptor-binding site, the overall distribution of evolutionary rates along the structure looked mostly random to the naked eye.

We thus found a geometrical, distance-based constraint on hemagglutinin evolution: Sites evolve the faster the closer they lie towards the receptor-binding region. Next, we wanted to evaluate how our distance metric performed relative to or in combination with other possible predictors of hemagglutinin evolution. One additional possible constraint on sequence divergence is a bias in the human immune system. This bias, generally referred to as antigenicity, describes the extent to which the human immune does a better job attacking one region of a protein compared to another. For influenza hemagglutinin H3, there exists a list of canonical, historical epitope sites that is commonly considered to represent this bias [4]. However, these sites were not primarily defined based on actual immunological data, and they have not been revisited since the late 1990s even though much more experimental data is now available (see Discussion for details on the history of the historical epitope sites). Therefore, before we could generate a combined evolutionary model, we considered it essential to re-derive the antigenic groups entirely from currently available immunological data.

Comparing historical epitope groups to experimental immune epitope data

We asked whether there is experimental evidence that the human immune system displays a bias towards particular regions of hemagglutinin; such a bias would traditionally be called an immune epitope. We obtained all available human B cell epitopes for H3 hemagglutinin from the Influenza Research Database (IRD). For comparison, we initially also considered all available non-human B cell epitope data in our analysis, because non-human data have traditionally been part of the data considered for epitope definition in influenza (see e.g. [10]). In the database, there were two types of B cell epitopes available for hemagglutinin: linear and non-linear. For humans, there were 31 separate epitope entries consisting of 26 non-linear and 5 linear. For non-humans, there were 134 available epitope entries consisting of 47 non-linear and 87 linear. The non-linear epitopes were provided as site numbers, so that mapping them onto the protein structure was trivial. By contrast, all linear epitopes were given as short peptide sequences with fewer than 40 amino acids per peptide. We mapped the linear epitopes onto the protein structure by jointly aligning the peptide sequences and the complete set of 3854 protein sequences used for evolutionary-rate calculations. Each linear epitope aligned without gaps to a particular segment of the full protein. The 5 linear human epitopes turned out to be a subset of the 87 non-human linear epitopes, and we therefore dropped the linear human epitopes from further analysis.

For both linear and non-linear epitope data, we counted for each site in hemagglutinin how often it appeared in each epitope data set. We then compared these epitope counts to the historical epitope sites (Bush 1999, Ref. [4]). We first considered the human, non-linear epitope data. We found some overlap between the historical sites and the experimental non-linear epitope counts (Fig. 2). Each of the four largest peaks (i.e., the four regions with strongest experimental evidence of belonging to epitopes) was at least partially captured by the historical epitope groups. The majority of historical epitopes A and B were represented in the non-linear epitope data; A had 16 of 19 sites appear in the experimental data and epitope B had 13 of 22. By contrast, many of the historical sites had no experimental support among the available non-linear epitope data. Of 131 historical epitope sites, only 52 appeared at least once in the experimental data set. Historical epitopes C, D, and E were particularly poorly represented as a fraction of their size, with only 8 of 27, 13 of 41, and 2 of 22 sites having experimental support, respectively. Among the discrepancies between historical and experimental immune epitopes, an important deviation came from the HA2 chain. By historical convention, no sites from HA2 were included among the hemagglutinin epitopes; as a result, there were a large number of sites appearing in the experimental set that were not defined as historical epitopes.

To understand to what extent the experimental non-linear epitope sites separated into distinct groups, we performed a clustering analysis (Fig. 3A). We considered all sites in the experimental epitope dataset as nodes of a graph, and we drew an edge between any two sites that appeared within the same accession number of the immune epitope database (IEDB). Then, we colored the nodes according to their classification in the historical epitope groups. We found that the non-linear experimental epitope data was able to partially reconstruct the historical epitope groups. In particular, historical epitopes A, B, and D clustered well, with only two sites from A and one site from D being completely disconnected from the rest of the epitope. By contrast, epitopes C and E did not at all recapitulate any experimental data. In addition, as previously shown in Fig. 2, there was a relatively large number of experimental epitope sites that were not accounted for by the historical epitope definition (Fig. 3A).

In addition to antibody connectivity (accession number clustering), given the structural nature of antibody neutralization, we expected that any correct grouping of epitope sites would display some ability to cluster in three dimensional space. Thus, we calculated the distance from each C_α (the α-Carbon atom in the polypeptide backbone) site in hemagglutinin to every other C_α. We then constructed a graph where we connected any two nodes in the non-linear experimental epitope data set with an edge if the two corresponding sites were less than 10 Å apart in the 3D structure (Fig. 3B). Again, we colored the nodes by their historical groupings. For these spatial clusters, the historical epitopes were poorly grouped within the non-linear epitope data. All historical epitopes fell into at least two disjoint sets separated by sites not belonging to the same epitope, and no single visible spatial cluster corresponded to a single historical epitope. We concluded that the historical epitope definitions largely failed to spatially cluster within the sites for which we had experimental immune epitope data.

We next considered the non-human, non-linear epitopes. We found that there was no clear connection between the non-human non-linear epitope data and historical epitopes groups (Fig. S1); in addition, several of the accessions listed individual sites, which thus did not provide any information about epitope groupings. We also considered experimental non-human linear epitopes. By contrast to the non-linear epitopes, the linear epitope set covered nearly every site in the entire hemagglutinin protein (Fig. S2). However, the most represented sites in the experimental epitope dataset (near site 100) were almost completely missed by the historical epitope sites. Furthermore, there was a substantial portion of experimental sites near the N-terminal region of HA1 that was completely absent from the historical sites. When clustering the data by shared accession numbers, we found that the experimental linear epitope data did not at all resemble historical epitope groups (Fig S3). Moreover, since the experimental linear epitopes covered almost every site in HA, it was not clear that they represented any particular immune-system bias. In fact, we expect that the experimental practice of expressing short linear peptides and testing them against antibody binding will generally produce many false positives, sites that are included in the peptide but not actually bound by an antibody.

Because neither the linear nor the non-linear non-human epitope sites appeared particularly informative, and because in general it is not clear that non-human immune data are relevant to human epitope grouping, we disregarded all non-human epitope data for the remainder of this study. We thus assumed that the non-linear human B cell epitopes represent the true immune epitope sites. We acknowledge that at least some of the sites included among the linear epitopes should likely be included among the true immune epitopes. However, these sites would first have to be verified by non-linear mapping.

Regrouping epitope sites with experimental data

Even though the historical epitopes were able to partially reconstruct the experimental epitope clusters, a simple visual inspection suggested more natural groups than those used in the historical set (Fig. 3A). Thus, we re-grouped the experimental human non-linear epitopes into the most obvious possible groupings (Fig. 3C and Table 1). The non-linear epitope data clustered most naturally into four distinct immune epitope regions, which we referred to as 1–4, to distinguish them from the historical epitopes A–E. Two of the four regions (experimental regions 2 and 3, respectively) were generally very similar to epitopes A and B in the historical definitions. One of the two remaining regrouped epitopes (experimental region 4) was vaguely similar to the D historical epitope with many sites added. The last epitope (experimental region 1) supported by experimental non-linear data was virtually nonexistent in all previous epitope groups. It had a few sites that were previously classified in the C epitope, and it added a large number of sites from the HA2 chain of hemagglutinin. Finally, there were eight sites that had at least one count in the non-linear epitopes and that could not be easily clustered with the other sites (Table 1).

We performed the same spatial clustering with our newly defined epitope groups as we had previously done for the historical epitope sites. By contrast to the historical epitopes, we found that our epitope groups, which were defined simply by antibody-clustered sites, partitioned almost perfectly into spatial clusters (Fig. 3D). Of the four groups we defined, all but one (experimental region 2) was spatially connected. In addition, experimental region 2 had only a single spatial disconnection. Our data set may be missing a single epitope site that would flip the relevant portion of the graph and connect the two disconnected sections. In addition, the spatial graph suggests how to resolve the ungrouped sites (drawn in black) in Fig. 3C. Of these eight sites, which grouped into one pair and two triplets in Fig. 3C, two likely belong to Epitope 2, two likely belong to epitope 3, and three may belong to either epitope 2 or 3. Notably, while our redefined epitopes clustered well in the 3D structure, they looked disconnected and arbitrarily chosen when plotted along the linear chain (Fig. S4).

Finally, we mapped our epitope groups onto the 3D crystal structure (Fig. 4), using the same color scheme as used in Figure 3. As expected, there was a clear spatial distribution of sites. Moreover, Epitopes 2–4 fell into the apical domain of HA, directly adjacent to the sialic acid-binding region. Only Epitope 1, which clearly separated from 2–4 in the clustering analysis, was located in the stem of HA.

Developing a predictive model of hemagglutinin evolution

Our ultimate goal in this work was to develop a predictive model of hemagglutinin evolution, a model that would use biophysical and/or biochemical properties of the protein to infer sites which are likely going to experience either positive or purifying selection pressure. What should such a model look like? Clearly, even in the complete absence of any host immune pressure not all sites in hemagglutinin are expected to evolve equally. In particular, several recent works have shown that site-specific evolutionary variation is partially predicted by a site's solvent exposure and/or number of residue-residue contacts in the 3D structure [19, 20, 25–30]. This relationship between protein structure and evolutionary conservation likely reflects the requirement for proper and stable protein folding: Mutations at buried sites or sites with many contacts are more likely to disrupt the protein's conformation [30] or thermodynamic stability [32]. In addition, there may be functional constraints on site evolution. For example, regions in proteins involved in protein–protein interactions or enzymatic reactions are frequently more conserved than other regions [27, 33, 34]. Likewise, for a viral surface protein, we expect that functionally important regions (regions with higher fitness consequences) in the protein that are targeted by antibodies will evolve more rapidly, to facilitate immune escape. And indeed, our results from the previous subsections have shown that (i) the neutralizing antibody epitopes in the IEDB preferentially target sites near the receptor-binding region in hemagglutinin and (ii) proximity to the receptor-binding region is a good predictor of evolutionary variation.

We therefore evaluated how proximity to the receptor-binding region performed as a predictor of dN/dS in comparison to the previously proposed predictors relative solvent accessibility (RSA) and weighted contact number (WCN). We found that among these three quantities, proximity to the sialic acid-binding region was the strongest predictor, explaining 16% of the variation in dN/dS (Pearson $r = 0.41$, $P < 10^{-15}$, see also Figs. 5 and S5). RSA and WCN explained

14% and 6% of the variation in dN/dS, respectively ($r = 0.37$, $P < 10^{-15}$ and $r = 0.25$, $P = 7 \times 10^{-9}$). Proximity to the sialic acid-binding region and RSA were virtually uncorrelated ($r = 0.08$, $P = 0.09$) while RSA and WCN correlated strongly ($r = -0.64$, $P < 10^{-15}$). These results suggested that proximity to the sialic acid-binding region and RSA should be used jointly in a predictive model. Importantly, proximity alone turned out to be the single strongest independent predictor of evolutionary rate currently known, outperforming not only RSA and WCN but also numerous other predictors previously considered [28].

We next asked to what extent epitope groups could predict evolutionary variation. We considered both the historical epitope groups (Bush 1999) and our experimentally derived epitopes 1–4, defined in the previous subsection. Because a site's epitope status is a categorical variable, we calculated variance explained as the coefficient of determination (R^2) in a linear model with dN/dS as the response variable and epitope status as the predictor variable. We found that experimental epitopes explained 15% of the variation in dN/dS, comparable to RSA and proximity. In comparison, the historical epitopes alone explained nearly 18% of the variation in dN/dS, outperforming all other individual predictor variables considered here (Fig. 5 and Table 2). However, as discussed in the previous subsection, the available experimental data suggest that not all of the historical sites may be actual immune epitope sites. Therefore, we suspected that some of the predictive power of historical sites was due to these sites simply being solvent-exposed sites near the receptor-binding region. We similarly wondered to what extent the predictive power of the experimental epitope sites was attributable to the same cause, since, in fact, both historical and experimental epitope sites showed comparable enrichment in sites near the sialic acid-binding region and in solvent-exposed sites (Fig. S6). Thus, we analyzed how the variance explained increased as we combined epitope sites (experimental or historical) with either RSA or proximity or both.

We found that epitope status, under either definition (experimental/historical), led to increased predictive power of the model when combined with either RSA or proximity (Fig. 5). However, a model consisting of just the two predictors RSA and proximity, not including any information about epitope status of any sites, performed even better than any of the other one- or two-predictor models, explaining 32% of the variation in dN/dS (Fig. 5). Adding epitope status to this best-performing two-predictor model resulted in only minor improvement, from 32% to 34% variance explained in the case of experimental epitopes and from 32% to 37% variance explained in the case of historical epitope sites (Fig. 5 and Table 2).

The geometrical constraints RSA and proximity explained more variance in dN/dS than did epitope sites, but were they also better at predicting sites of interest? Because dN/dS can measure purifying as well as positive diversifying selection, the percent variance in dN/dS that a model explains may not necessarily accurately reflect how useful that model is in predicting specific sites, e.g. sites under positive selection. For example, one could imagine a scenario in which a model does exceptionally well on sites under purifying selection ($dN/dS \ll 1$) but fails entirely on sites under positive selection ($dN/dS > 1$). Such a model might explain a large proportion of variance but be considered less useful than a model that overall predicts less variation in dN/dS but accurately pinpoints site under positive selection. Therefore, we wondered whether epitope sites might do a poor job predicting background purifying selection but might still be useful in predicting sites with $dN/dS > 1$. We found, to the contrary, that neither the historical nor the experimental epitope sites could reliably predict sites with

$dN/dS > 1$, alone or in combination with RSA (Fig. S7A–D). Proximity to the receptor-binding site, on the other hand, correctly predicted four sites with $dN/dS > 1$, even in the absence of any other predictors. Notably, all models we considered here were robust to cross-validation. The cross-validated residual standard error was virtually unchanged from its non-cross-validated value in all cases (Table 2). Because proximity clearly identified four points with high dN/dS, we also verified that the proximity–dN/dS correlation was not caused just by these four points. We removed from our data set the four points that had both predicted and observed $dN/dS > 1$, and found that a significant proximity–dN/dS correlation remained nonetheless ($r = 0.17$, $p = 0.00001$).

Finally, we compared the predictions from the geometrical model of hemagglutinin evolution to results from a recent study of antigenic cluster transitions; that study found seven sites near the receptor-binding region which were critical for cluster transitions according to hemagglutinin inhibition (HI) assays with ferret antisera [21]. The sites identified in Ref. [21] were 145, 155, 156, 158, 159, 189, and 193. For comparison, our geometric model (with predictors RSA and 1/Distance) predicted none of these sites to be under positive selection. Sites predicted to have $dN/dS > 1$ were instead 96, 137, 138, 143, 222, 223, 225, and 226. Moreover, out of the seven sites from Ref. [21], only one (site 145) had an observed dN/dS significantly above 1. By contrast, four of the eight sites predicted under the geometric model to have $dN/dS > 1$ did indeed have dN/dS significantly above 1. Thus, the sites that determine the major antigenic changes in the virus did not at all overlap with the sites expected and observed to be under the greatest evolutionary pressure. When investigating the location of these sites in detail, we found that all of the sites we predicted to have $dN/dS > 1$ were located just basal to the receptor-binding site, whereas nearly all of the sites from [21] (with the exception of 145, the site with $dN/dS > 1$) were located on the apical side of the receptor-binding site (Fig. S8).

In summary, we have found that two simple geometric measures of a site's location in the 3D protein structure, solvent exposure and proximity to the receptor-binding region, jointly outperformed, by a wide margin, any previously considered predictor of evolutionary variation in hemagglutinin, including immune epitope sites. In fact, the vast majority of the variation in evolutionary rate that was explained by the historical epitope sites was likely due to these sites simply being located near the receptor-binding region on the surface of the protein. However, historical epitope sites, in combination with solvent exposure and proximity, had some residual explanatory power beyond even a three-predictor model that combined the two geometric measures with experimental immune-epitope data. We suspect that this residual explanatory power reflects the sequence-based origin of the historical epitope sites. To our knowledge, the historical epitope sites were at least partially identified by observed sequence variation, so that, to some extent, these sites are simply the sites that have been observed to evolve rapidly in hemagglutinin.

Discussion

We have conducted a thorough analysis of the determinants of site-specific hemagglutinin evolution. Most importantly, we have found that host immune bias (as measured by experimentally determined immune epitopes) accounts for a very small but significant portion of the evolutionary pressure on influenza hemagglutinin. By contrast, a simple geometric measure, receptor-binding proximity, is the single largest constraint on hemagglutinin evolution we have identified, and the only quantity that can predict sites with $dN/dS > 1$. We have shown that the historical epitope definitions overlap only partially with experimentally determined immune-epitope data, and we have defined new epitopes based on the experimentally available data. These experimentally supported epitopes cluster in protein tertiary structure space. Finally, we have shown that a simple linear model containing three predictors, solvent accessibility, proximity to the receptor-binding region, and immune-epitope sites, explains nearly 35% of the evolutionary rate variation in hemagglutinin H3.

History of epitopes in hemagglutinin H3

Efforts to define immune epitope sites in H3 hemagglutinin go back to the early 1980's [10]. Initially, epitope sites were identified primarily by speculating about the chemical neutrality of amino acid substitutions between 1968 (the year H3N2 emerged) and 1977, though some limited experimental data on neutralizing antibodies was also considered [10, 11]. In 1981, the initial four epitope groups were defined by non-neutrality (amino-acid substitutions that the authors believed changed the chemical nature of the side chain) and relative location, and given the names A through D [10]. Since that original study in 1981, the names and general locations of H3 epitopes have remained largely unchanged [4, 16]. The sites were slightly revised in 1987 by the same authors and an additional epitope named E was defined [11]. From that point forward until 1999 there were essentially no revisions to the codified epitope sites. In addition, while epitopes have since been redefined by adding or removing sites, no other epitope groups have been added [3,16,18]; epitopes are still named A–E. In 1999, the epitopes were redefined by more than doubling the total number of sites and expanding all of the epitope groups [4]. At that time, the redefinition consisted almost entirely of adding sites; very few sites were eliminated from the epitope groups. Although this set of sites and their groupings remain by far the most cited epitope sites, it is not particularly clear what data justified this definition. Moreover, when the immune epitope database (IEDB) summarized the publicly available data for flu in 2007, it only included one experimental B cell epitope in humans (Table 2 in [35]). Although there were a substantial number of putative T cell epitopes in the database, *a priori* there is no reason to expect a T cell epitope to show preference to hemagglutinin as opposed to any other influenza protein; yet it is known that several other influenza proteins show almost no sites under positive selection. Moreover, it is known that the B cell response plays the biggest role is maintaining immunological memory to flu, and thus it is the most important arm of the adaptive immune system for influenza to avoid.

The historic H3 epitope sites have played a crucial role in molecular evolution research. Since 1987, an enormous number of methods have been developed to analyze the molecular evolution of proteins, and specifically, to identify positive selection. The vast majority of these methods have either used hemagglutinin for testing, have used the epitopes for validation, or have at some point been applied to hemagglutinin. Most importantly, in all this work, the epitope definitions have been considered fixed. Most investigators simply conclude that their methods work as expected because they recover some portion of the epitope sites. Yet virtually all of these studies identify many sites that appear to be positively selected but are not part of the epitopes. Likewise, there is no single study that has ever found all of the epitope sites to be important. Even if the identified sites from all available studies were aggregated, we would likely not find every site among the historical epitopes in that aggregated set of sites.

Implications of historical epitope groups for current research

Given all of this research activity, it seems that the meaning of an immune epitope has been muddled. Strictly speaking, an immune epitope is a site to which the immune system reacts. There is no *a priori* reason why an immune epitope needs to be under positive selection, needs to be a site that has some number or chemical type of amino acid substitutions, or needs to be predictive of influenza whole–genome or hemagglutinin specific sequence cluster transitions. Yet, from the beginning of the effort to define hemagglutinin immune epitopes, such features have been used to identify epitope sites, resulting in a set of sites that may not accurately reflect the sites against which the human immune system produces antibodies.

Ironically, this methodological confusion has actually been largely beneficial to the field of hemagglutinin evolution. As our data indicate, if the field had been strict in its pursuit of immune epitopes sites, it would have been much harder to produce predictive models with those sites, in particular given that experimental data on non-linear epitopes have been sparse until very recently. By contrast, the historical epitope sites have been used quite successfully in several predictive models of the episodic nature of influenza sequence evolution. In fact, in our analysis, historical epitopes displayed the highest amount of variance explained among all individual predictors (Fig. 5). We argue here that the success of historical epitope sites likely stems from the fact that they were produced by disparate analyses each of which accounted for a different portion of the evolutionary pressures on hemagglutinin. Of course, it is important to realize that some of this success is likely the result of circular reasoning, since the sites themselves were identified at least partially from sequence analysis that included the clustered, episodic nature of influenza hemagglutinin sequence evolution.

Despite the success of historical epitope groups, they only predict about 18% of the evolutionary rate-variation of hemagglutinin. Moreover, since many of these sites likely are not true immune epitopes (and therefore not host dependent), one might ask which features of the historical epitope sites make them good predictors. We suspect that they perform well primarily because they are a collection of solvent-exposed sites near the sialic acid-binding region (see also Fig. S6). We had shown previously that sites within 8 Å of the sialic acid-binding site are enriched in sites under positive selection, compared to the rest of the protein [20]. A similar result was found in the original paper by Bush et al. [4]. However, the related metric of distance from the sialic acid-binding site has not previously been considered as a predictor of evolution in hemagglutinin. Furthermore, before 1999, most researchers thought the opposite should be true; that receptor-binding sites should have depressed evolutionary rates [4]. Even today the field seems split on the matter [21]. As we have shown here, the inverse of the distance from sialic acid is the single *best* independent predictor of hemagglutinin evolution; by itself this distance metric can account for 16% of evolutionary rate-variation. Moreover, by combining this one metric with another to control for solvent exposure, we can account for more than a third of the evolutionary rate variation in hemagglutinin. For reference, this number is larger than the variation one could predict by collecting and analyzing all of the hemagglutinin sequences that infect birds (another group of animals with large numbers of natural flu infections), and using those rates to predict human influenza hemagglutinin evolutionary rates [20].

In this context, it is important to note that the IDB has limitations; not all existing (not to mention all possible) immunological data have been added. Further, the extent to which certain epitopes (e.g., stalk epitopes) have been mapped may be more reflective of a bias in research interests among influenza researchers than a bias in the human immune system. Therefore, in our re-derivation of epitope groupings, we may be missing immune epitope sites or be incorrectly grouping the ones that we have. Our analysis of epitope sites will likely have to be redone as more data become available. However, we expect that as more non-linear data become available, they will broadly follow the trend observed in the linear epitope data, that is, the more antibodies are mapped, the more sites in the hemagglutinin protein appear in at least one mapping, until virtually every site in the entire hemagglutinin protein is represented. Under this scenario, the ability to predict evolution from immunological data would become worse, not better, as more data are accumulated.

Geometric constraints likely dominate adaptive evolution in hemagglutinin

Why do geometric constraints (solvent exposure and proximity to receptor-binding site) do such a good job predicting hemagglutinin evolutionary rates? Hemagglutinin falls into a class of proteins known collectively as viral spike glycoproteins (GP). In general, the function of these proteins is to bind a host receptor to initiate and carry out uptake or fusion with the host cell. Therefore, a priori one might expect that the receptor-binding region would be the most conserved part of the protein, since binding is required for viral entry. Yet in hemagglutinin sites near the binding region are the most variable in the entire protein. There are at least two possible models that might explain this observation. First, in terms of host immune evasion, antibodies that bind near the receptor-binding region may be the most inhibitory, and hence mutations in this region the most effective in allowing immune escape. Viral spike GPs have a surface that is both critical for viral survival and is sufficiently long lived that a host immune response is easily generated against it. There are likely many other viral protein surfaces that are comparatively less important or sufficiently short lived during a conformational change that antibody neutralization is impractical. Thus, the virions that survive to the next generation are those with substantial variation at the surface or surfaces with high fitness consequences and a long half-life in vivo. Evolutionary variation at surfaces with low or no fitness consequences, or at short-lived surfaces, should behave mostly like neutral variation and hence appear as random noise, not producing a consistent signal of positive selection. Second, according to the avidity modulation model of Hensley et al. [23], it is possible that antibody inhibition is not overcome by escaping the antibody directly. Rather, a single or a few relatively rare mutations may increase the avidity of hemagglutinin for its receptor so as to out-compete partial antibody inhibition. Subsequently, once the partial inhibition is overcome in a competent host, passage to an incompetent how allows genetic drift to bring the avidity back down to baseline. Both of these models are reasonable under our analyses in this study.

We also need to consider that actual epitope sites, i.e., sites towards which the immune system has a bias, may not be that important for the evolution of viruses. An epitope is simply a part of a viral protein to which the immune system reacts. Therefore, it represents a host-centered biological bias. The virus may experience stronger selection at regions with high fitness consequences but that generate a relatively moderate host response compared to other sites with low fitness consequences that generate a relatively strong host response. For example, if an antibody binds to the receptor-binding region of hemagglutinin, the influenza virus must adapt to survive; by contrast, if an antibody binds away from the receptor-binding region, the influenza virus may not be neutralized by that antibody and hence may not experience much selection pressure to adapt. For these reasons, we expect that the geometric constraints we have identified here will be more useful in future modeling work than the experimental epitope groups we have defined. Moreover, we expect that similar geometrical constraints will exist in other viral spike glycoproteins, and in particular in other hemagglutinin variants.

Remarkably, the sites we found that experienced the most positive selection showed minimal overlap with the sites found to be minimally sufficient for explaining the major antigenic transitions in H3N2, as determined by HI assays with ferret antisera [21]. While both groups of sites lie near the sialic-acid binding region, the vast majority of positively selected sites are located basally to sialic acid whereas sites identified by HI assays lie predominantly on the apical side (Fig. S8). This finding suggests that HI assays and positive selection analyses reflect distinct biological mechanisms. For example, HI assays might not accurately reflect selection pressures *in vivo*. Alternatively, HI assays may correctly identify mutations that lead to antigenic cluster transitions whereas positive selection analyses may identify sites that mediate avidity [23] or antigenic drift within a cluster. In future work, disentangling the different mechanisms reflected by HI assays and by positive-selection analyses will likely be crucial for improved prediction of HA evolution and of optimal vaccine strains.

Materials and Methods

Obtaining influenza data and preparing sequences

All of the data we analyzed were taken from the Influenza Research Database (IRD) [36]. The IRD provides experimental immune epitope data curated from the data available in the Immune Epitope Database (IEDB) [37].

We used sequences that had been collected since the 1991–1992 flu season. Any season before the 1991–1992 season had an insufficient number of sequences to contribute much to the selection analysis. The sequences were filtered to remove redundant sequences and laboratory strains. The sequences were then aligned with MAFFT [38]. Since it is known that there have been no insertions or deletions since the introduction of the H3N2 strain, we imposed a strict opening penalty and removed any sequences that had intragenic gaps. In addition, we manually curated the entire set to remove any sequence that obviously did not align to the vast majority of the set; in total the final step only removed about 10 sequences from the final set of 3854 sequences. For the subsequent evolutionary rate calculations, we built a tree with FastTree 2.0 [39].

Computing evolutionary rates and relative solvent accessibilities

To compute evolutionary rates, we used a fixed effects likelihood (FEL) approach with the MG94 substitution model [24,40,41]. We used the FEL provided with the HyPhy package [24]. For the full setup see the linked GitHub repository (`https://github.com/wilkelab/influenza_HA_evolution`). As is the case for all FEL models, an independent evolutionary rate is fit to each site using only the data from that column of the alignment. Because our data set consisted of nearly 4000 sequences, almost every site in our alignment had a statistically significant posterior probability of being either positively or negatively selected after adjusting via the false discovery rate (FDR) method. As shown in Figure S7, all evolutionary rates fall into a range between $dN/dS = 0$ and $dN/dS = 4$.

We computed RSA values as described previously [28]. Briefly, we used DSSP [42] to compute the solvent accessibility of each amino acid in the hemagglutinin protein. Then, we used the maximum solvent accessibilities [43] for each amino acid to normalized the solvent accessibilities to relative values between 0 and 1. We found that RSA calculated in the trimeric state produced better predictions than RSA calculated in the monomeric state. Thus, we used multimeric RSA in all models in this study. Both multimeric and monomeric RSA are included in the supplementary data.

Mapping experimental immune epitopes

There were two broad categories of epitope data available in the epitope database. One was non-linear epitopes and the other was linear epitopes. In addition, the IRD splits epitopes into B and T cell epitopes and also into the various species whose immune system was tested. Here, we were interested in antibody-driven immune escape, and hence we focused on B cell epitopes.

For B cell epitopes, every linear human epitope was also recognized by the immune system of many other species. Therefore, we decided to split the analysis into linear and non-linear groups. All of the non-linear epitope data came in the form of amino acid plus site number. As a result, we could map the epitopes by simply counting the number of times each site in the protein was hit and the co-occurrences of sites hit by the same antibody.

Mapping linear epitopes was slightly more complicated. We started with the short sequence fragments; each of the fragments was between 5 and 40 amino acids in length. We tried initially to map the fragments onto the emblematic A/Aichi/2/1968 sequence, but it turned out the epitope fragments were actually generated from disparate strains along the H3N2 lineage. The differences between the original founder strain and the fragments meant that we could not accurately map the vast majority of short peptides onto a single sequence. Instead, we took the entire curated and pre-aligned set of 3854 sequences that we used in the evolutionary rate calculations. We then aligned the fragments with MAFFT using a very strong opening penalty of 10. We visually checked to be sure all of the 87 fragments aligned reasonably well to the full H3 alignment. Then, as with the non-linear epitopes, we counted the number of times each site was hit and the co-occurrences of sites hit by the same antibody.

To calculate the distance maps, we used an H3 protein structure (PDB ID: 4FNK) available in the Protein Databank [44,45]. We first cleaned the protein structure using PyMOL [46]. We then used the Bio.PDB module of biopython to compute distances on the protein structure. In a similar fashion to the co-occurrence maps above, we used every site that appeared in at least one epitope. We then calculated the distance from every experimental epitope site to every other experimental epitope site. If a site was within 10 Å of another epitope site, then an edge was drawn between them using the igraph package [47] in the statistical programming language R [48].

Evolutionary rate-distance correlations

To create the structural heat map of correlations shown in Fig. 1B, we first needed to calculate the correlations between evolutionary rates and pairwise distances, calculated in turn for each location in the protein structure as the reference point for the distance calculations. Conceptually, we can think of this analysis as overlaying a grid on the entire protein structure, where we first calculate the distance to various grid points from every C_α in the entire protein, and then compute the correlation between the set of distances to the sites on the grid and the evolutionary rate at those sites. In practice, we calculated the distance from each C_α to every other C_α. We then colored each residue by the correlation obtained between evolutionary rates and all distances to its C_α.

Statistical analysis and data availability

All statistical analyses were performed using R [48]. We built the linear models with both the `lm()` and `glm()` functions. For cross validation, we used the `cv.glm()` function within the boot package. Residual standard error values were computed by taking the square root of the delta value from `cv.glm()`. With the exception of graph visualizations, all figures in this manuscript were created using ggplot2 [49].

A complete data set including evolutionary rates, epitope assignments, RSA, and proximity to the receptor-binding site is available as Table S1. Raw data and analysis scripts are available at
`https://github.com/wilkelab/influenza_HA_evolution`. In the repository, we have included all human H3 sequences from the 1991–1992 season to present combined into a single alignment. We have cleaned the combined data to only include sequences with canonical bases, non-repetitive sequences, and we have hand filtered the data to ensure all included sequences align appropriately to the 566 known amino acid sites. In addition, we have built a tree and visually verified that there were no outlying sequences on the tree for the combined set.

Technical considerations for analysis

The site-wise numbering for the H3 hemagglutinin protein reflects the numbering of the mature protein; this numbering scheme requires the removal of the first 16 amino acids in the full-length gene. Thus, for protein numbering purposes, site number 1 is actually the 17th codon in full-length gene numbering. The complete length of the H3 hemagglutinin gene is 566 sites while the total length of the protein is 550 sites. It is important to point out that the mature H3 protein has two chains (HA1 and HA2) that are produced by cutting the presursor (HA0) protein between sites 329 and 330 in protein numbering. In addition, as a result of cloning and experimental diffraction limitations, most (or likely all) hemagglutinin structures do not include some portion of the first or last few amino acids of either chain of the mature protein, and crystallographers always remove the C-terminal transmembrane span from HA2. For example, the structure we used (PDBID: 4FNK) in this study does not include the first 8 amino acids of HA1, the last 3 amino acids of HA1, or the last 48 amino acids of HA2. As a result, HA1 includes sites 9–326 and HA2 includes sites 330–502. Table S1 lists the gene sequence from one of the three original H3N2 (Hong Kong flu) hemagglutinin (A/Aichi/2/1968), the gene numbering, the protein numbering, the numbering of one H3N2 crystal structure, historical immune epitope sites from 1981, 1987 and 1999, and every calculated parameter used (and many others than were not used) in this study. In general, the most common epitope definitions in use today are those employed by Bush et. al 1999 [4]. Throughout this work, we refer to the Bush et. al 1999 epitopes as the“historical epitope sites”.

Acknowledgments

We would like to thank Jesse Bloom for helpful comments on this manuscript and Robin Bush for providing us with a complete list of the historical epitope groupings.

References

1. Pybus OG, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. Nature Rev Genet 10: 540–550.

2. Bhatt S, Holmes EC, Pybus OG (2011) The genomic rate of molecular adaptation of the human influenza A virus. Mol Biol Evol 28: 2443–2451.

3. Luksza M, Lassig M (2014) A predictive fitness model for influenza evolution. Nature 507: 57–61.

4. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. Science 286: 1921–1925.

5. Koelle K, Cobey S, Grenfell B, Pascual M (2006) Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. Science 314: 1898–1903.

6. 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.

7. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, et al. (2014) Integrating influenza antigenic dynamics with molecular evolution. eLife 3: e01914.

8. Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ (2006) Long intervals of stasis punctuated by burst of positive selection in the seasonal evolution of influenza a virus. Biology Direct 1: 34.

9. Vijaykrishna D, Smith GJD, Pybus OG, Zhu H, Bhatt S, et al. (2011) Long-term evolution and transmission dynamics of swine influenza A virus. Nature 1473: 519–522.

10. Wiley DC, Wilson IA, Skehel JJ (1981) Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature 289: 373–378.

11. Wiley DC, Skehel JJ (1987) The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Ann Rev Biochem 56: 365–394.

12. Bush RM, Fitch WM, Bender CA, Cox NJ (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 16: 1457–1465.

13. Skehel JJ, Wiley DC (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Ann Rev Biochem 69: 531–569.

14. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, et al. (2004) Mapping the antigenic and genectic evolution of influenza virus. Science 205: 371–375.

15. Suzuki Y (2006) Natural selection on the influenza virus genome. Mol Biol Evol 23: 1902–1911.

16. Shih AC, Hsiao T, Ho M, Li W (2007) Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution. Proc Natl Acad Sci USA 104: 6283–6288.

17. Tamuri AU, dos Reis M, Hay AJ, Goldstein RA (2009) Identifying changes in selective constraints: Host shifts in influenza. PLoS Comput Biol 5: e1000564.

18. Pan K, Deem MW (2011) Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. J Roy Soc Interface 8: 1644-1653.

19. Meyer AG, Wilke CO (2013) Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30: 36–44.

20. Meyer AG, Dawson ET, Wilke CO (2013) Cross-species comparison of site-specific evolutionary-rate variation in influenza hemagglutinin. Phil Trans R Soc B 368: 20120334.

21. Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GCM, et al. (2013) Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science 342: 976–979.

22. Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. eLife 3: e03568.

23. Hensley SE, Das SR, Bailey AL, Schmidt LM, Hickman HD, et al. (2009) Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science 326: 734–736.

24. Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenetics. Bioinformatics 21: 676–679.

25. Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291: 177–196.

26. Bustamante CD, Townsend JP, Hartl DL (2000) Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. Mol Biol Evol 17: 301–308.

27. Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol 26: 2387–2395.

28. Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, et al. (2014) Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. J Mol Evol 79: 130–142.

29. Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, et al. (2014) Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. Mol Biol Evol 31: 135–139.

30. Huang TT, Marcos ML, Hwang JK, Echave J (2014) A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol Biol 14: 78.

31. Sikosek T, Chan HS (2014) Biophysics of protein evolution and evolutionary protein biophysics. J Royal Soc Interface 11: 20140419.

32. Echave J, Jackson EL, Wilke CO (2014) Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. bioRxivorg : http://dx.doi.org/10.1101/009423.

33. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257: 342-358.

34. Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314: 1938–1941.

35. Bui H, Peters B, Assarsson E, Mbawuike I, Sette A (2007) Ab and T cell epitopes of influenza A virus, knowledge and oppurtunities. Proc Natl Acad Sci USA 104: 246–251.

36. Squires RB, Noronha J, Hunt V, García-Sastre A, Macken C, et al. (2012). Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza and Other Respiratory Viruses, DOI:10.1111/j.1750-2659.2011.00331.x.

37. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, et al. (2010) The immune epitope database 2.0. Nucleic Acids Res 38: D854–62.

38. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772–780.

39. Price MN, Dehal PS, Arkin AP (2009) FastTree 2 – approximately maximum-likelihood trees for large alignments. PLOS ONE 5: e9490.

40. Yang Z (2006) Computational Molecular Evolution. Oxford University Press.

41. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715–724.

42. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.

43. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO (2013) Maximum allowed solvent accessibilites of residues in proteins. PLOS ONE 8: e80635.

44. Berman HM (2008) The Protein Data Bank: a historical perspective. Acta Crystallographica Section A: Foundations of Crystallography A64: 88–95.

45. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.

46. Schrödinger, LLC (2010) The PyMOL molecular graphics system, version 1.3r1.

47. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems: 1695.

48. Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5: 299–314.

49. Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer New York. URL `http://had.co.nz/ggplot2/book`.

Tables

Table 1. Groups of experimental non-linear epitope sites. We defined epitope groups on the basis of which sites co-occurred within the same IEDB accession number. See also Fig. 3C and Table S1. Sites are numbered according to their position in the mature protein.

Epitope group	Sites
1	34, 36, 53, 54, 70, 292, 295, 305, 307, 334, 363, 364, 365, 366, 379, 380, 382, 383, 384, 386, 387, 390, 391, 393, 394, 395, 397, 398, 401, 403, 404, 405, 499
2	121, 122, 123, 124, 126, 131, 133, 135, 136, 137, 138, 140, 142, 143, 144, 145, 146
3	155, 156, 157, 158, 159, 160, 187, 188, 189, 190, 191, 192, 194, 196, 223, 256
4	114, 115, 147, 148, 149, 150, 151, 152, 153, 154, 161, 162, 169, 170, 171, 172, 173, 174, 175, 176, 204, 205, 206, 208, 209, 210, 211, 212, 235, 238, 241, 242, 243
N/A	82, 83, 222, 225, 275, 276, 278

Table 2. Predictive performance of each linear model considered. R^2 is the proportion of variation in dN/dS explained by the specified model. RSE is the residual standard error of the linear model. cvRSE$_{10}$ is the cross validated residual standard error calculated by 10-fold cross validation. cvRSE$_{\mathrm{loo}}$ is the cross validated residual standard error calculated by leave-one-out cross validation.

Predictors in the linear model	R^2	RSE	cvRSE$_{10}$	cvRSE$_{\mathrm{loo}}$
RSA	0.14	0.41	0.41	0.41
Experimental epitopes	0.15	0.41	0.42	0.42
1 / Distance	0.16	0.40	0.41	0.41
Bush 1999	0.18	0.40	0.41	0.41
RSA + Experimental epitopes	0.23	0.39	0.41	0.40
RSA + Bush 1999	0.24	0.39	0.39	0.39
1 / Distance + Experimental epitopes	0.23	0.39	0.40	0.40
1 / Distance + Bush 1999	0.28	0.38	0.39	0.39
RSA + 1 / Distance	0.32	0.37	0.37	0.37
RSA + 1 / Distance + Experimental epitopes	0.34	0.36	0.39	0.38
RSA + 1 / Distance + Bush 1999	0.37	0.35	0.37	0.37

Figures

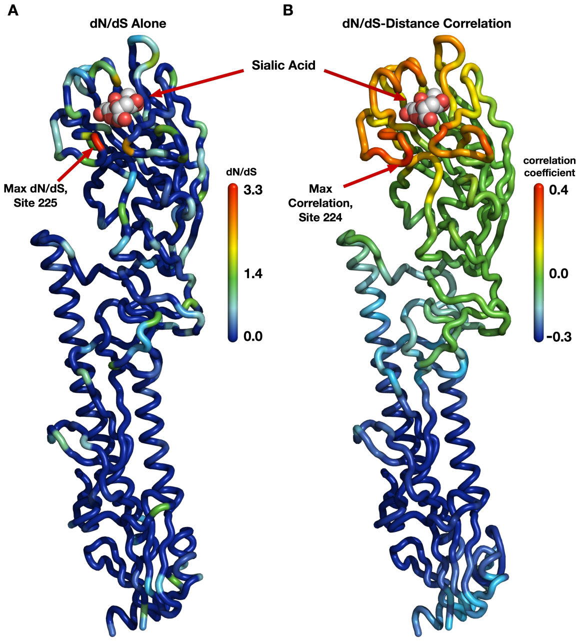

Figure 1. Evolutionary-rate variation along the hemagglutinin structure. (A) Each site in the protein structure is colored according to its evolutionary rate dN/dS. Hot colors represent high dN/dS (positive selection) while cool colors represent low dN/dS (purifying selection). (B) Each site in the protein structure is colored according to the dN/dS–distance correlation obtained when distances are calculated relative to that site. Hot colors represent positive correlations while cool colors represent negative correlations. Thus, distances from sites that are redder are better positive predictors of the evolutionary rates in the protein than are distances from bluer sites; distances from blue sites are actually anti-correlated with evolutionary rate. Distances from sites that are colored green have essentially no predictive ability.

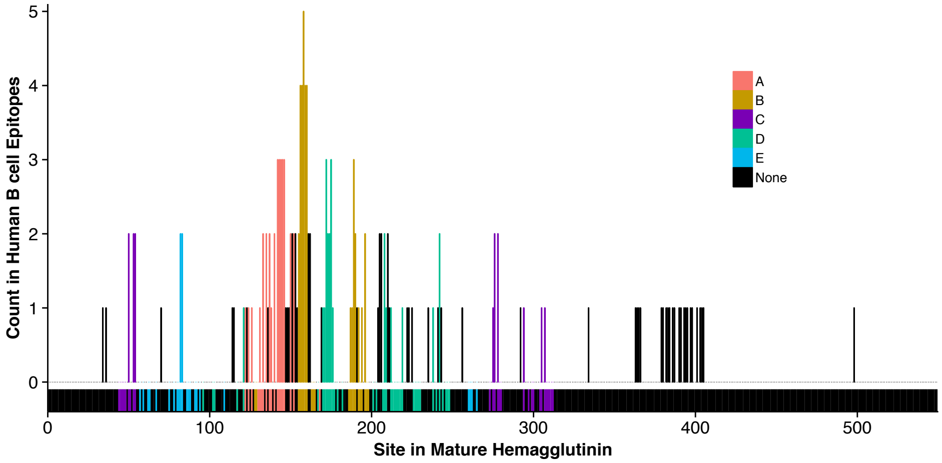

Figure 2. Experimental non-linear epitope site counts, colored by historical epitope groupings. The heights of individual bars indicate how often each site in the H3 hemagglutinin protein appears in an experimental non-linear epitope set, and the color of each bar indicates the site's historical epitope assignment according to Bush et. al 1999 [4]. Sites that do not appear in the experimental set are shown with a count of zero. The rug underneath the $y = 0$ line contains all sites and visualizes the exact loation of the historical epitope sites.

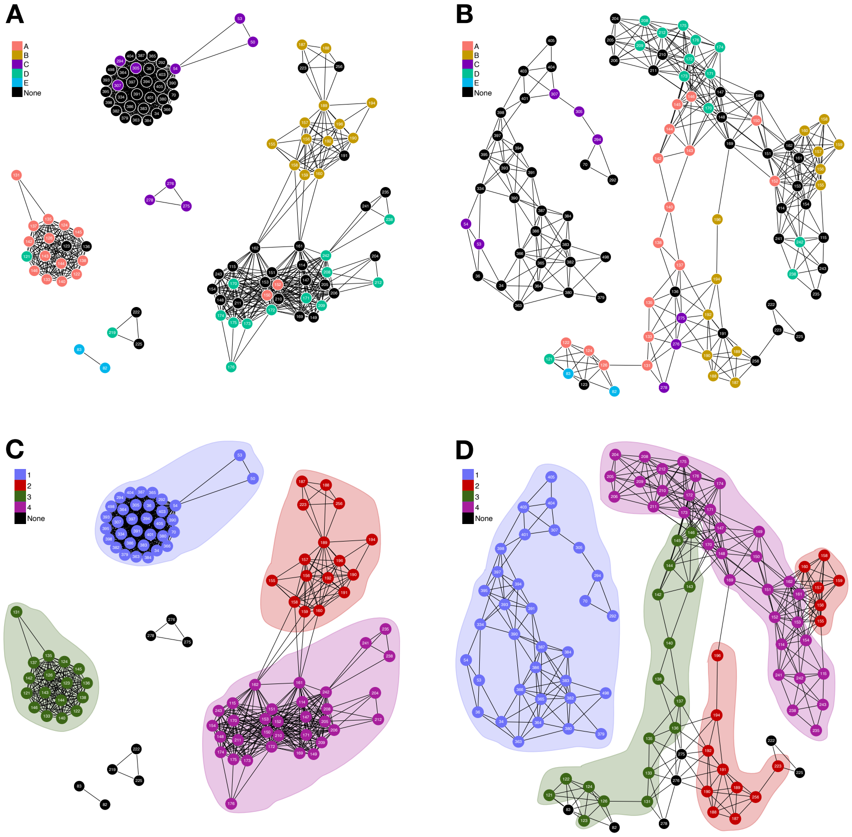

Figure 3. Clustering of experimental non-linear epitope sites by IEDB accession number and by physical proximity in the 3D structure. (A) Non-linear epitope sites clustered by IEDB accession number and colored according to the historical epitope definition [4]. Each node represents a site that appears at least once in the experimental non-linear epitope data set. Two nodes are connected by an edge if they are bound by the same antibody in the Immune Epitope Database (i.e., have the same IEDB accession number). The historical epitope definitions do not correspond well to the experimentally observed clustering. (B) Non-linear epitope sites clustered by physical proximity in the 3D structure and colored according to the historical epitope definition. The same nodes as in (A) are now connected by an edge if they are within 10 Å of each other in the three dimensional structure. The historical epitope groupings do not appear clustered in 3D space. (C) Sites are clustered as in A but colored according to the most natural grouping based on antibody clustering. Colored areas are drawn to highlight the distinct clusters. (D) Site are clustered as in (B) but colored as in (C). Parts (C) and (D) show that the experimental non-linear epitope data support the expectation that immune epitopes should group together spatially. All the sites that group by binding the same set of antibodies are also spatially connected. The only exception is the red epitope (#2), which is split into two groups. Also, there is one site that appears once in the data, but is not connected to any other sites; as a result, it could not be displayed.

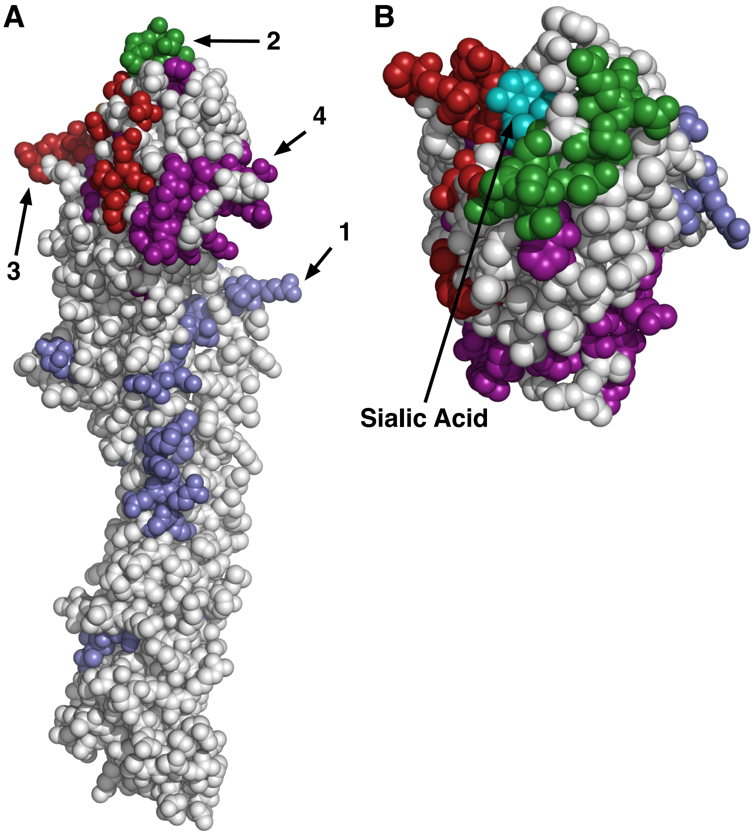

Figure 4. Location of experimental, non-linear epitopes in the 3D structure of hemagglutinin. (A) Side perspective of hemagglutinin. Non-linear epitope sites are colored according to their group assignment, as defined in Fig. 3. (B) Top perspective of hemagglutinin, with epitope sites highlighted. The orange moiety is sialic acid: the human receptor for influenza.

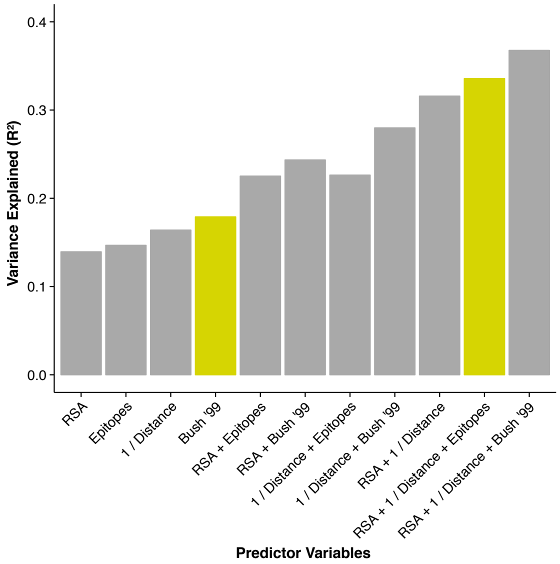

Figure 5. Proportion of variance in dN/dS explained by different linear models. The height of each bar represents the coefficient of determination (R^2) for a linear model consisting of the stated predictor variables. The historical epitope sites from Bush 1999 [4] (yellow bar on the left) are the single best predictor of evolutionary rate variation. However, a model using three predictors that each have a clear biophysical meaning (solvent exposure, proximity to receptor-binding region, non-linear epitope status) explains almost twice the variation in dN/dS (yellow bar on the right).

Supplementary Tables

Table S1: Complete data set including evolutionary rates, solvent accessibilities, proximities to the receptor-binding region, and epitope status for all sites.

Supplementary Figures

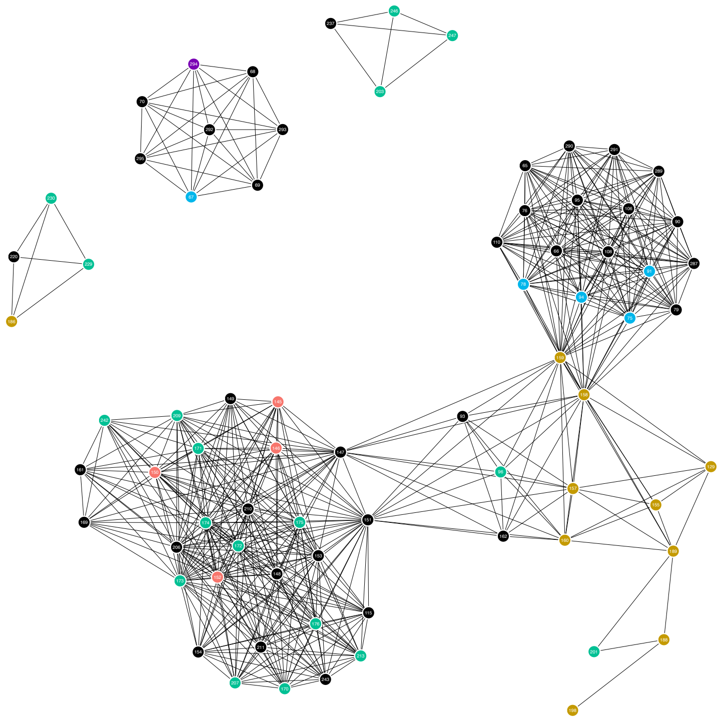

Figure S1: Clustering of experimental non-human non-linear epitope sites by IEDB accession number. Each node represents a site that appears at least once and is connected to another site in the experimental non-linear epitope data set. Nodes are colored according to the historical epitope definition [4]. Two nodes are connected by an edge if they are both part of the same IEDB accession number. The historical epitope definitions do not correspond well to the experimentally observed clustering. Also, there are 13 sites that appear once in the data but are not connected to any other sites; as a result, they could not be displayed.

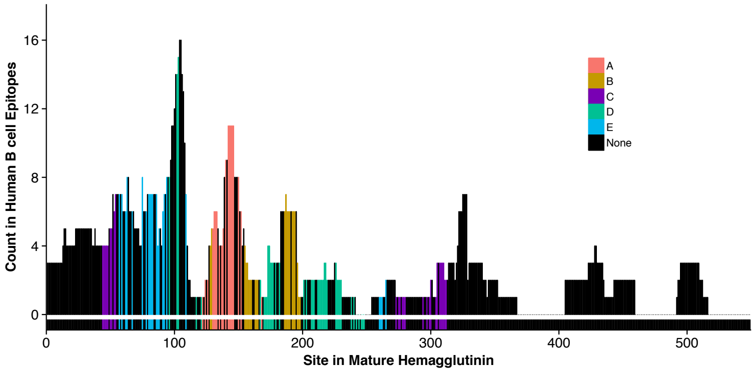

Figure S2: Experimental non-human linear epitope site counts, colored by historical epitope groupings. The heights of individual bars indicate how often each site in the H3 hemagglutinin protein appears in an experimental linear epitope set, and the color of each bar indicates the site's historical epitope assignment according to Bush et. al 1999 [4]. Sites that do not appear in the experimental set are shown with a count of zero. The rug underneath the $y = 0$ line contains all sites and visualizes the exact loation of the historical epitope sites.

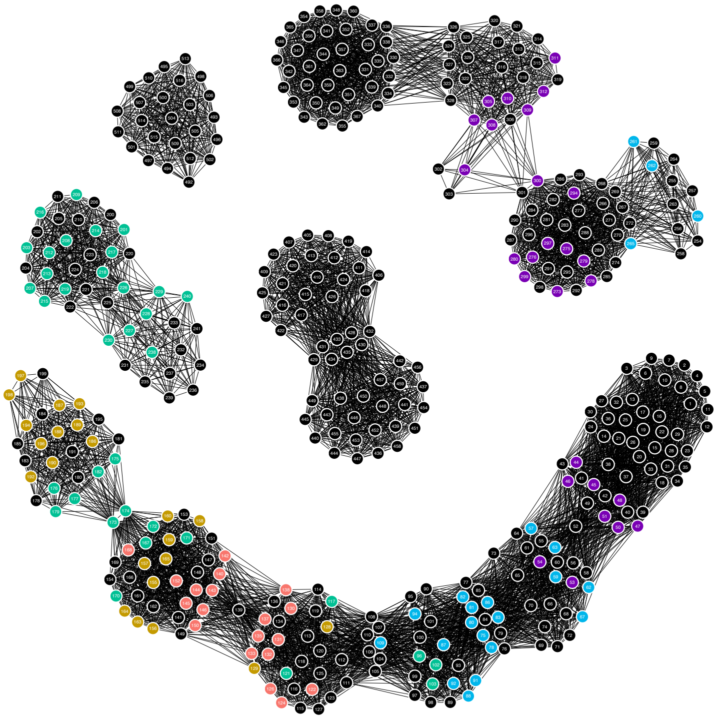

Figure S3: Clustering of experimental non-human linear epitope sites by IEDB accession number. Each node represents a site that appears at least once and is connected to another site in the experimental linear epitope data set. Nodes are colored according to the historical epitope definition [4]. Two nodes are connected by an edge if they are both part of the same IEDB accession number. The historical epitope definitions do not correspond well to the experimentally observed clustering.

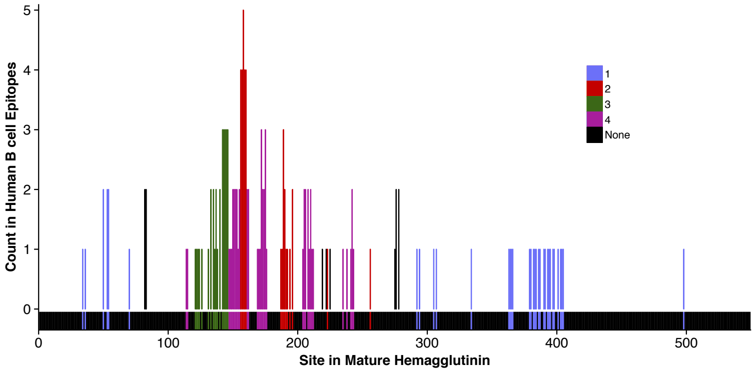

Figure S4: Experimental human non-linear epitope site counts, colored by our proposed, experimentally-based epitope groupings. The heights of individual bars indicate how often each site in the H3 hemagglutinin protein appears in an experimental non-linear epitope set, and the color of each bar indicates the site's epitope assignment. Sites that do not appear in the experimental set are shown with a count of zero. The experimentally-based epitopes are not clustered linearly along the sequence, but instead fall into a non-contiguous spatial arrangement.

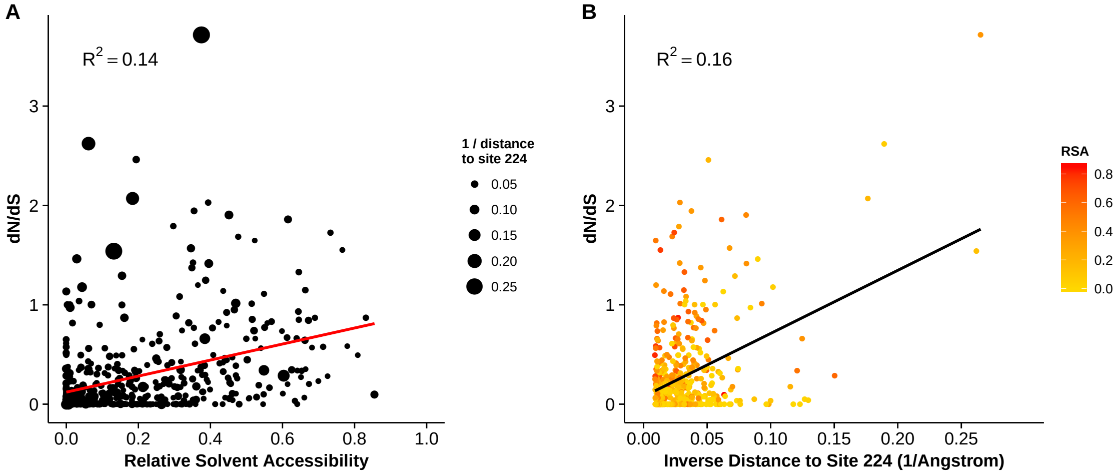

Figure S5: Dependence of dN/dS on solvent exposure and proximity to the receptor-binding region. (A) dN/dS vs. RSA. The size of the dots represents 1/Distance. (B) dN/dS vs. 1/Distance. The coloring of the dots represents RSA. The distance to the sialic acid-binding region is the single strongest quantitative predictor of evolutionary rate ratio in hemagglutinin.

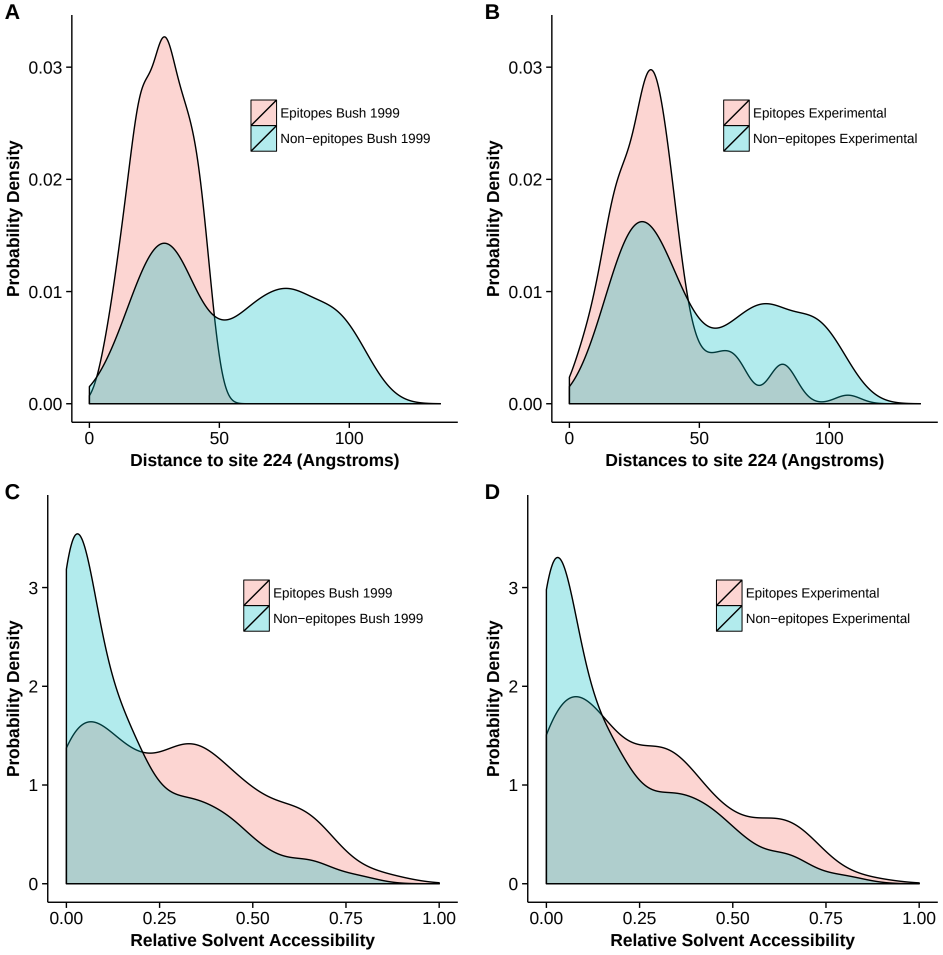

Figure S6: Distance to receptor-binding site and solvent exposure for epitope and non-epitope sites. (A) Distribution of distances to residue 224, for historical epitope and non-epitope sites. (B) Distribution of distances to residue 224, for experimental non-linear epitope and non-epitope sites. (C) Distribution of relative solvent accessibilities, for historical epitope and non-epitope sites. (D) Distribution of relative solvent accessibilities, for experimental non-linear epitope and non-epitope sites. Under both historical and experimental epitope definitions, epitope sites are closer to the sialic acid-binding region and have higher RSA than non-epitope sites.

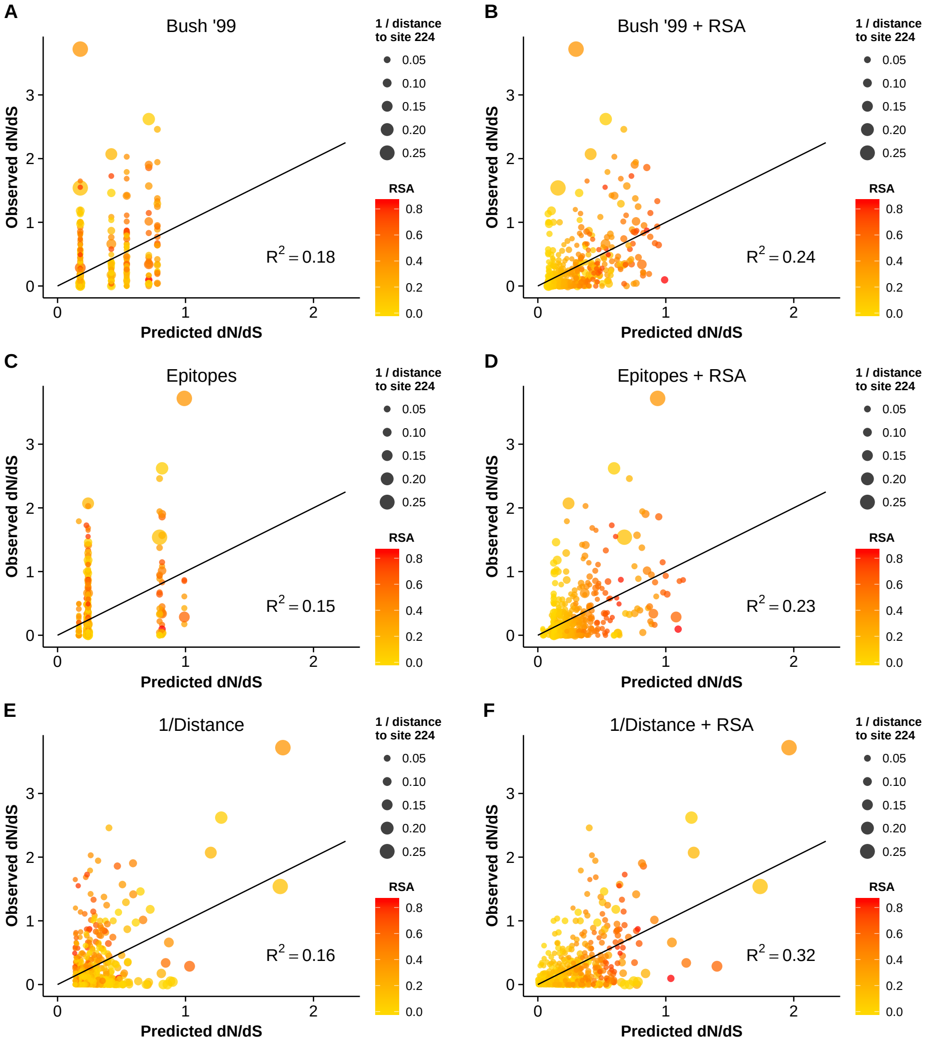

Figure S7: Observed dN/dS vs. predicted dN/dS for different predictive linear models. (A) Only epitope status according to the historical definition is used as predictor variable. (B) Historical epitope sites and RSA are used as predictor variables. (C) Only epitope status according to the experimental non-linear epitope data is used as predictor variable. (D) Experimental epitope sites and RSA are used as predictor variables. (E) Only proximity to the sialic acid-binding region (measured as 1/Distance to Residue 224) is used as predictor variable. (F) Proximity and RSA are used as predictor variables. Individual sites with $dN/dS > 1$ are predicted correctly only if the linear model includes the 1/Distance predictor. However, in all cases, adding the RSA predictor significantly improves the model predictions.

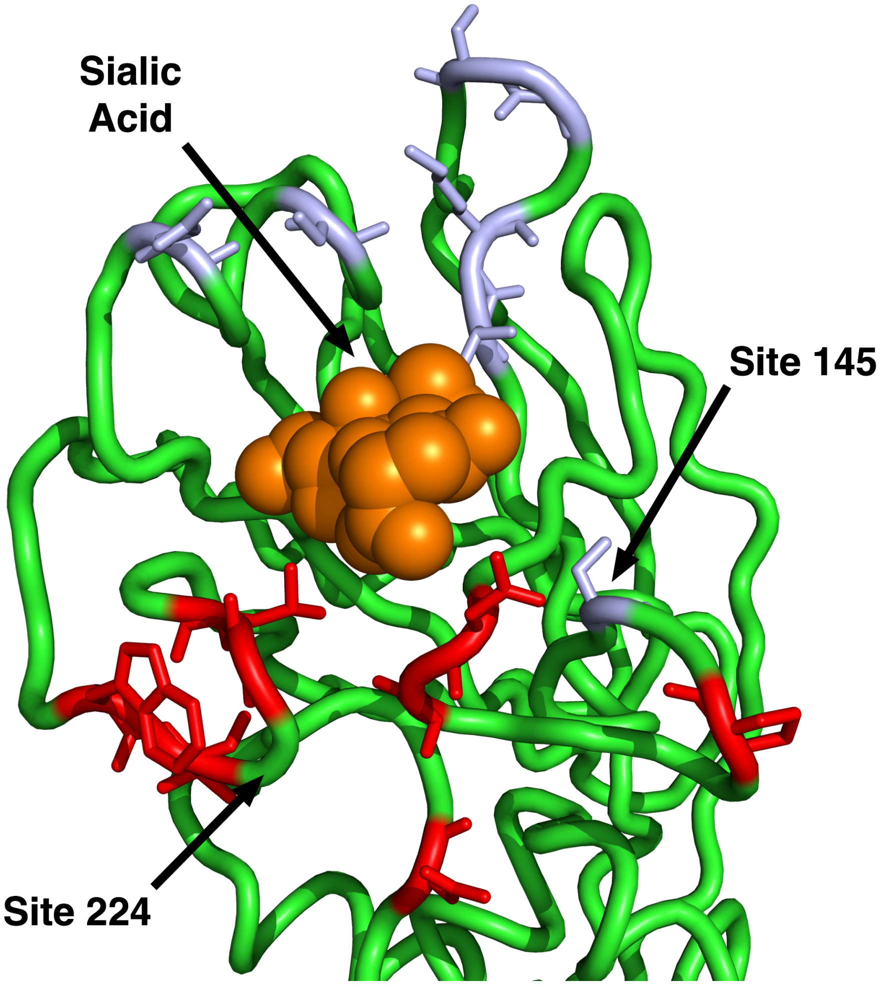

Figure S8: Sites identified by Koel et al. 2013 and those predicted to have $dN/dS > 1$. The sites shown in purple are those identified by Koel et al. 2013 [21] to be critical for antigenic cluster transitions. Only one of these sites has a dN/dS significantly above one, site 145. The sites shown in red are those that our geometrical model predicts to have $dN/dS > 1$. (Half of those sites have observed $dN/dS > 1$.) Note that our model predicts only sites on the basal side of sialic acid to be under positive selection, since our reference point for proximity is site 224. Site 145, the only purple site under positive selection, is also the only purple site on the basal side of sialic acid.