Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin

Austin G. Meyer, Eric T. Dawson and Claus O. Wilke

Section of Integrative Biology, Institute for Cellular and Molecular Biology, and Center for Computational Biology and Bioinformatics, The University of Texas, Austin, Austin, TX 78731, USA

We investigate the causes of site-specific evolutionary-rate variation in influenza haemagglutinin (HA) between human and avian influenza, for subtypes H1, H3, and H5. By calculating the evolutionary-rate ratio, \( \omega = \frac{dN}{dS} \) as a function of a residue's solvent accessibility in the three-dimensional protein structure, we show that solvent accessibility has a significant but relatively modest effect on site-specific rate variation. By comparing rates within HA subtypes among host species, we derive an upper limit to the amount of variation that can be explained by structural constraints of any kind. Protein structure explains only 20–40% of the variation in \( \omega \). Finally, by comparing \( \omega \) at sites near the sialic-acid-binding region to \( \omega \) at other sites, we show that \( \omega \) near the sialic-acid-binding region is significantly elevated in both human and avian influenza, with the exception of avian H5. We conclude that protein structure, HA subtype, and host biology all impose distinct selection pressures on sites in influenza HA.

1. Introduction

Viral proteins are highly variable at the sequence level; they accumulate amino acid substitutions at a rapid pace [1,2]. Yet their structures tend to be fairly conserved. Highly variable surface regions notwithstanding, most viral proteins need to maintain a specific structure to carry out their function in the viral life cycle [3]. The generally accepted picture is that sites in the protein core maintain the overall protein structure and are, therefore, most conserved. Sites on the surface are less critical to the protein structure and hence more free to vary, for example in response to selection pressures imposed by immune response. This view is based on the finding, replicated in widely differing organisms and using many different techniques, that, on average, sequence variability increases the closer a site is located towards the surface of a protein [4–13]. More specifically, in influenza, exposed sites in haemagglutinin (HA) and neuraminidase have been found to evolve faster than buried sites in these proteins [14,15].

Thus, prior work has clearly established that protein structure influences site variability. What is less clear, however, is the magnitude of this effect. Is knowing a site is buried sufficient to predict that the site will be evolutionarily conserved, or are other factors stronger driving forces for site-specific evolutionary rates? And similarly, will homologous sites in related but distinct viral strains evolve at similar rates, or do the nature of the viral strain and the infected host organism impose stronger influences on site-specific evolutionary rates than the location of a site in the protein structure?

Here, we address these questions for influenza HA. We compare per-site sequence evolution for two different host species (human and avian) and three HA subtypes (H1, H3 and H5), and ask the following questions: (i) To what extent is rate variation determined by the location of a site in the structure, as measured by the sites' relative solvent accessibility (RSA)? (ii) To what extent is rate variation conserved within HA subtypes among viruses infecting different host species? (iii) Are \( \omega = \frac{dN}{dS} \) ratios elevated near the active site (the...
sialic-acid-binding region, SABR) of HA? We find that protein structure, HA subtype and host biology affect rate variation in influenza HA.

2. Material and methods

(a) Sequence preparation

We obtained sequences for HA subtypes H1, H3 and H5 for human and avian hosts from the Influenza Research Database [16]. Using the built-in curating tools of the database, we carefully selected subsets of sequences that corresponded as much as possible to well-defined and distinct viral populations. Sequences were curated within each host species depending on its subtype. In particular, for each combination of HA subtype and host species, we considered only sequences that could be linked to a specific neuraminidase subtype.

Human H1 sequences were obtained from H1N1 strains isolated between 1977 (after the Fort Dix outbreak) and 2008 (before the 2009 flu pandemic). H1N1 strains since 2009 are not direct descendants of H1N1 strains before 2009 and thus were excluded. We found 2057 distinct H1 sequences. Human H3 sequences were obtained from H3N2 strains isolated between 1968 and 2012. We found 8315 distinct sequences. Human H5 sequences were obtained from H5N1 sequences without date restriction. We found 297 distinct sequences.

Avian sequences were curated by subtype with no restrictions placed on the date range; full datasets from FluDB of H1N1, H3N2 and H5N1 sequences were used. We found 106, 115 and 2684 distinct sequences, respectively.

To align sequences and map them to the HA structure, we downloaded a reference structure from the protein data bank (PDB). We used the structure with PDB identifier 1rd8, which is a structure of the H1 subtype [17]. All nucleotide sequences were translated into amino acid sequences. Each distinct set of sequences (human H1, human H3 and so on) was then aligned to the reference amino acid sequence from the PDB file, using the MUSCLE sequence-alignment tool with default settings [18]. After alignment, gaps that were introduced relative to the sequence in the PDB file were removed, and the amino acids were reverted to their nucleotide codons. For alignments that contained more than 200 sequences, we selected a random subset of 200 sequences for further analysis. This reduction in alignment sizes was necessary because maximum-likelihood (ML) fitting of codon-evolution models becomes prohibitively slow for much larger alignments.

For each alignment, we generated a phylogenetic tree by ML using the program RAxML [19]. This tree was used as the base tree for the evolutionary-rate analyses described in the later section. We used the GTRCAT substitution approximation available in RAxML; this approximation was chosen to make computing a phylogenetic tree computationally tractable with the large number of sequences used here. Similarly, the multi-threading option was used in the RAxMLHPC version to speed up the computation. The following command was used to generate the phylogenetic tree:

\[
\text{raxmlHPC -T 2 -m GTRCAT -s input_file -n tree}
\]

To assess the phylogenetic relationship between human and avian alignments, we also constructed combined human/avian alignments for H1, H3 and H5, and reconstructed trees for the combined alignments. For H1, we found nearly complete phylogenetic separation between human and avian sequences. All but three avian and all but two human sequences formed a separate clade each. The remaining five sequences (introduced from classical swine lineage) were closer to each other than to either of the two clades. For H3, human and avian sequences were mostly separated. However, 14 avian sequences fell into the human clade (12 of those avian sequences grouped together as a single clade, introduced from the triple-reactassortant H3N2 swine lineage). For H5, we found no phylogenetic separation. Human and avian sequences grouped together throughout the entire phylogeny. Human and avian H5 viruses are not distinct lineages, because human H5 is not transmitted effectively from human to human. All phylogenetic trees are provided as electronic supplementary materials.

(b) Relative solvent accessibility determination and binning

We calculated RSA as described in Meyer & Wilke [15]. In brief, solvent accessibilities (SAs) were calculated with the program DSSP [20] and then normalized [9]. The sequence data were then subdivided into eight evenly spaced bins according to the RSA of their sites in the protein structure, as described in Scherrer et al. [13].

(c) Evolutionary-rate determination

We calculated site-specific evolutionary rates using two approaches that integrate sequence data and protein structure, a random-effects likelihood (REL) and a fixed-effects likelihood (FEL) approach. Both approaches were implemented in the phylogenetic-modelling language HyPhy [21], and our HyPhy scripts are provided as part of the electronic supplementary materials.

The REL approach was previously described in Meyer & Wilke [15]. In brief, we used the standard Goldman–Yang codon-evolution model (GY94 [22]), but made the evolutionary-rate ratio \( \omega = dN/dS \), the branch length \( t \) and the transition to transversion ratio \( \kappa \) linear functions in RSA. We express their RSA dependence as

\[
\omega(\text{RSA}) = \omega_0 + \omega_1 \times \text{RSA} + \omega_2,
\]

\[
t(\text{RSA}) = t_0 + t_1 \times \text{RSA} + t_2 \quad \text{and} \quad \kappa(\text{RSA}) = \kappa_0 + \kappa_1 \times \text{RSA} + \kappa_2.
\]

(2.1)

(2.2)

(2.3)

(The parameters \( t \) and \( \kappa \) are nuisance parameters for the purpose of our analysis here, but need to be fit with RSA dependence for optimum results, as discussed in the recent studies [13,15].) Further, \( \omega_0 \) and \( \omega_0 \) are random effects, drawn from discrete distributions with a finite set of categories. Specifically, the distribution of \( \omega_0 \) is described by pairs of values \((\omega_0, p_0^k)\), such that \( \Pr(\omega_0 = \omega_0^k) = p_0^k \), where \( k = 1, \ldots, D_\omega \), \( p_0^k \geq 0 \), \( \sum_k p_0^k = 1 \). Similarly, the distribution of \( \omega_0 \) is described by pairs of values \((\omega_0, p_0^k)\), where \( k = 1, \ldots, D_\omega \). The parameter \( D_\omega \) determines the number of \( \omega \)-slopes in our model, and the parameter \( D_\omega \) determines the number of \( \omega \)-intercepts. All other parameters \((t_0, t_1, \kappa_0 \) and \( \kappa_0 \) were fixed effects.

The infinitesimal matrix generator \( Q \) for the GY94 model has the usual form (for \( i \neq j \))

\[
Q_{ij} = \begin{cases}
0, & \text{if more than one nucleotides changes} \\
\pi_{ij}, & \text{if synonymous transition} \\
\kappa \pi_{ij}, & \text{if non-synonymous transition} \end{cases}
\]

where \( \pi_{ij} \) is the frequency of codon \( j \), the indices \( i \) and \( j \) run over all 61 sense codons, and \( \kappa \) and \( \omega \) are RSA-dependent, as stated above. The codon frequencies \( \pi_{ij} \) were calculated from the entire sequence alignment using the F3×4 model. All other parameters were fitted by ML.

To identify the optimal number of slopes (\( D_\omega \)) and intercepts (\( D_\omega \)) for the \( \omega \)-RSA dependence, we used the Akaike information criterion (AIC; [23,24]). We fit between zero and five slopes, and between one and five intercepts in all pairwise combinations. We defined the best-fitting model as the one with the minimum AIC value, as described in Meyer & Wilke[15].

Site-specific evolutionary rates were calculated by first calculating posterior probabilities for each site and rate category, using
the empirical Bayes approach [25], and then averaging over all rate categories at each site, as described in Meyer & Wilke [15].

The FEL model is built on top of the REL model, and can be considered a structure-aware modification of existing FEL approaches [26]. After determining the best-fitting REL model, we re-fit a GY94 model separately at each site. In this fit, the variables \( \alpha \) and \( t \) are not re-fitted to the data, but instead set at each site to the values predicted from the REL model. Thus, the FEL approach fits only a single parameter for each site, the evolutionary-rate ratio \( \omega \).

All data files and analysis scripts are provided as electronic supplementary materials.

### 3. Results

(a) Estimating site-specific evolutionary rates in a structural context

We constructed separate sequence alignments for human and avian H1, human and avian H3, and human and avian H5 (six separate alignments; see §2 for a detailed description). We then calculated site-specific evolutionary-rate ratios, \( \omega = \frac{dN}{dS} \), for each alignment. We carried out these calculations in a structure-aware framework; our goal was to determine the relative importance of protein structure and other factors for evolutionary-rate variation among sites.

There are two alternative approaches to calculating per-site evolutionary rates, the REL method and FEL method. Under the REL method, a finite set of \( \omega \) categories is fit jointly to all sites in the alignment. Sites are then probabilistically assigned to categories using an empirical Bayes approach [25,27]. To obtain distinct \( \omega \) values for each site, we calculate averages over all \( \omega \) categories, weighted by each site’s posterior probability to belong to each category [15]. Under the FEL method, a distinct \( \omega \) value is fit separately to each codon column in the alignment. Existing REL and FEL methods tend to have comparable performance, with either one being preferable in certain applications [26].

We recently introduced a structure-aware REL model, which calculates \( \omega \) taking into account the RSA of individual residues. RSA is a measure of how close to the surface a residue is located in the folded, three-dimensional protein structure. A residue with an RSA of 0 is located entirely inside the protein structure; it has no contact with water. A residue with an RSA of 1, on the other hand, is completely exposed. Such residues tend to occur in variable surface loops. Intermediate RSA values correspond to partially buried residues.

In the structurally aware REL model, \( \omega \) is written as a linear function of RSA, \( \omega = \omega_0 \times \text{RSA} + \omega_b \), and both \( \omega_0 \) and \( \omega_b \) are random variables drawn from discrete distributions with a fixed set of categories (see [15] and §2). The optimal number of slope \( (\omega_0) \) and intercept \( (\omega_b) \) categories is determined by minimizing the AIC. This model generally produces a better fit than a comparable model without a structural component. Typical numbers of categories needed for both slopes and intercepts fall between two and four [15].

We fit this model to all six HA alignments, and found that the RSA dependence was significant in all cases except one (human H5). The optimal number of slopes and intercepts varied by alignment (table 1).

We also compared \( \omega \) estimates predicted by the best REL model with RSA dependence with those predicted by the best REL model without RSA dependence, and found that estimates were generally very similar, with correlation coefficients above 0.9 (table 1). We would like to emphasize, however, that this similarity only arose when we averaged \( \omega \) over rate categories, as described in Meyer & Wilke [15]. Thus, even though using a model with RSA dependence is preferable in terms of model fit, a model without RSA dependence can perform almost as well, as long as \( \omega \) is averaged over rate categories.

Next, we assessed the reliability of per-site \( \omega \) ratios, by comparing REL estimates with FEL estimates. Such a comparison has not previously been done for the structure-aware REL model. We implemented a structure-aware FEL model (§2), fitted it to all alignments, and found that REL and FEL estimates were generally very similar (figure 1).

Correlation coefficients were typically above 0.9. FEL tended to produce slightly higher \( \omega \) estimates for large \( \omega \) values and slightly lower estimates for low \( \omega \) values. This finding is consistent with the prior observation of shrinkage in REL estimates relative to FEL estimates [26]. One downside of the REL approach is excessive shrinkage, if the alignment in small. This problem does not arise in the FEL approach. By contrast, in the FEL approach the number of sequences in the alignment imposes a lower limit on the smallest \( \omega \) values that can be estimated. All estimates below this limit are simply zero. The REL method, by contrast, assigns rates to these sites based on the

### Table 1. Comparison of RSA-dependent and RSA-independent model fits.

<table>
<thead>
<tr>
<th>host</th>
<th>HA type</th>
<th>slopes, intercepts</th>
<th>best overall</th>
<th>best no-RSA</th>
<th>best RSA</th>
<th>corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>H1</td>
<td>3, 2</td>
<td>0, 3</td>
<td>3, 2</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>H3</td>
<td>2, 3</td>
<td>0, 3</td>
<td>2, 3</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>H5</td>
<td>0, 3</td>
<td>0, 3</td>
<td>2, 3</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>avian</td>
<td>H1</td>
<td>2, 3</td>
<td>0, 4</td>
<td>2, 3</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>avian</td>
<td>H3</td>
<td>2, 2</td>
<td>0, 3</td>
<td>2, 2</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>avian</td>
<td>H5</td>
<td>4, 3</td>
<td>0, 4</td>
<td>4, 3</td>
<td>0.96</td>
<td></td>
</tr>
</tbody>
</table>

aNumber of slopes and intercepts for the overall best-fitting model.
bNumber of slopes and intercepts for the best-fitting model without RSA dependence.
cNumber of slopes and intercepts for the best-fitting model with RSA dependence.
dCorrelation (corr.) of log\( \omega \) between the best model with and without RSA. All correlations are highly significant (\( p < 10^{-10} \)).
density of sites with no non-synonymous variation relative to the density of sites with some non-synonymous variation. Because the REL and FEL methods provide comparable estimates and REL provides a positive for all sites, we used REL for the remainder of this paper. All results remained qualitatively unchanged when analyses were carried out using the FEL method.

(b) Elevated near the sialic-acid-binding region
When plotting per-site against RSA, we found that generally increased with RSA, as expected (figure 2).
However, the rate variation around this overall trend was large. varied over one to two orders of magnitude at all RSA values. We found the least rate variation in human and avian H1, and the most variation in human H5.
To assess how variation in $\omega$ correlated with protein function, we next compared $\omega$ of sites near the SABR with the $\omega$ of other sites. We considered sites to be near the SABR, if the corresponding residue had at least one atom within 8Å of the sialic-acid molecule in our reference PDB structure. There were 39 such sites. In the following, we will refer to these sites as SABR sites. The cutoff of 8Å is conservative, in the sense that we likely captured all sites relevant for sialic-acid binding as well as some sites that are not. Thus, even if the entire SABR was under positive selection, we would not expect all 39 SABR sites to have elevated $\omega$. However, we would expect this set of sites to be enriched for sites with elevated $\omega$ relative to other regions in the protein.

Visual inspection of figure 2 reveals that SABR sites have elevated $\omega$ for both human and avian viruses for most HA subtypes. We quantified this association by estimating how these sites as SABR sites. The cutoff of 8Å is conservative, in the sense that we likely captured all sites relevant for sialic-acid binding as well as some sites that are not. Thus, even if the entire SABR was under positive selection, we would not expect all 39 SABR sites to have elevated $\omega$. However, we would expect this set of sites to be enriched for sites with elevated $\omega$ relative to other regions in the protein.

We found systematic differences in H1 and H3 across host species; $\omega$ estimates for avian H1 and H3 were generally lower than the corresponding estimates for human H1 and H3 (figure 3). Human and avian H5, on the other hand, showed virtually no difference in $\omega$ on average. For SABR sites specifically, we found that their $\omega$ differences were mostly in agreement with the overall $\omega$ differences of the strains compared. We tested whether SABR sites were more likely to fall either above or below the lines of maximum covariation in figure 3, using Fisher’s exact test as before. For H1 and H3, we could not reject the null hypothesis of an odds ratio of 1. However, we did find that SABR sites evolved at a significantly elevated $\omega$ in human H5 compared with avian H5.

(c) Site-specific $\omega$ estimates vary substantially among host species and haemagglutinin subtypes We next assessed to what extent site-specific evolutionary-rate ratios were comparable across species. We found that while there was a clear trend towards similar $\omega$ across species, variation around this trend was substantial (figure 3).

Correlation coefficients fell between 0.49 and 0.6. Thus, between 24 and 36 per cent of the variation in $\omega$ for one host species was explained by the $\omega$ values of sequences infecting another host species. More specifically, in any given comparison, strong negative selection (low $\omega$) in one group did not guarantee strong negative selection in another group. Several of the most constrained sites in avian sequences had $\omega$ near or above 1 in human sequences (figure 3).

Despite the overall finding that conservation in one group does not guarantee conservation in the other group, most comparisons also showed clusters of sites with very similar (and typically low) $\omega$ values. Thus, there seems to be a core set of conserved sites that are shared among strains.

We found systematic differences in H1 and H3 across host species; $\omega$ estimates for avian H1 and H3 were generally lower than the corresponding estimates for human H1 and H3 (figure 3). Human and avian H5, on the other hand, showed virtually no difference in $\omega$ on average. For SABR sites specifically, we found that their $\omega$ differences were mostly in agreement with the overall $\omega$ differences of the strains compared. We tested whether SABR sites were more likely to fall either above or below the lines of maximum covariation in figure 3, using Fisher’s exact test as before. For H1 and H3, we could not reject the null hypothesis of an odds ratio of 1. However, we did find that SABR sites evolved at a significantly elevated $\omega$ in human H5 compared with avian H5.

(d) Differences in $\omega$ are moderately biased towards the protein core

Finally, we wanted to assess whether the largest differences in $\omega$ among groups associated with specific regions in the structure. To this end, we introduced a function $G$, defined as $G(1, 2) = |\log(\omega_1/\omega_2)|$, which measures the absolute difference in $\log\omega$ at a site. We plotted $G$ as a function of RSA to determine whether the largest differences occurred on the protein surface or in the core (figure 4).

We found a weak trend of declining $G$ with increasing RSA in H1 and H3, and the opposite in H5 (figure 4). Thus, for H1 and H3, sites in the core of the protein showed, on average, slightly larger differences in $\omega$ than sites near the surface. This finding suggests that $\omega$ differences among strains are not just caused by different antigenic pressure (which would exert itself on the surface of the protein), but also by some processes that act throughout, and possibly particularly towards the core of, the protein structure. The inverted trend for H5 may indicate that in these viruses, the difference between the avian and the human selective environment is indeed dominated by surface-bound processes (antibody binding and/or sialic-acid binding).

We mapped the SABR sites onto the G-RSA plots to determine whether they were contributing to an excess of buried sites with high $\omega$ in humans. Clearly, many of the SABR sites have relatively low RSA (figures 2 and 4). However,
no obvious trend emerged. SABR sites did not cause the negative correlation between $G$ and RSA observed for H1 and H3.

4. Discussion

We have found that protein structure makes a significant contribution to per-site evolutionary-rate variation in influenza HA. We have also found that sites near the SABR show elevated $\omega$ in most strains. Finally, we have found that site-specific rate variation among different influenza strains is comparable, but by no means identical. This finding emphasizes the diverse nature of strain and host-specific selection pressures throughout the HA protein.

The strength of the correlations among $\omega$ for different host species inform us about the extent to which protein structure constrains sequence evolution. If protein structure were entirely responsible for any rate variation, then these correlations should be near 100 per cent. By contrast, if protein structure had no influence on sequence variation, then these correlations should be near 0 per cent (to the extent that evolution has truly been restricted to only one host species). We found that the $\omega$ values of sequences for one host species explained between 20 and 40 per cent of the variation in $\omega$ for sequences of another host species (correlation coefficients fell between 0.49 and 0.6). Thus, at least 60 per cent, and possibly more, of the variation in $\omega$ is not explained by protein structure.

We found that all viral strains except avian H5 experienced elevated $\omega$ in the SABR. This finding likely reflects diversifying selection in this region, driven by antibody escape or HA-receptor interactions. In humans, antibodies are not generally thought to bind at the SABR, yet such antibodies do exist and may be sufficiently common to drive positive selection in this region [28]. Further, the SABR is small compared with the typical binding region of an antibody. Thus, mutations in the periphery of the receptor pocket can interfere with antibody binding without inhibiting sialic-acid binding [28]. Second, the positively selected mutants may not primarily be antibody-escape mutants, but rather mutants that change the avidity of HA-receptor interactions [29]. Such substitutions could be driven by sequential passage in immune and naive hosts, for example human adults and human children, where immune hosts select for increased avidity and naive hosts select for decreased avidity [29].

In our comparison of strains infecting different host species (figure 3), we found for H1 and H3 that evolutionary-rate ratios $\omega$ in the SABR were consistent with the overall differences in evolutionary rates among host type. For example, the comparison of avian to human H3 shows that avian H3 is overall more conserved than human H3, and this increased conservation is visible in similar magnitude near the SABR and elsewhere in the protein. Further, both the H1 and the H3 comparison show clear clusters of sites that are conserved in both strains but overall more conserved in one strain than in the other (i.e. clusters are shifted away from the dashed line representing equal $\omega$). Note that this result is not likely to be caused by mutation-rate differences, because $\omega = dN/dS$ is normalized for mutation rate. In combination, these findings suggest that there are systematic differences in selection pressures among human and avian influenza strains. These altered selection pressures seem to operate throughout the entire protein, not just in specific regions. Previous studies have shown that the SABR is more conserved in avian viruses than in human viruses [30]. Our results do not disagree with this previous observation, but they suggest that this observation may have been caused by selection pressures acting on the entire HA protein, rather than by selection pressures acting specifically on the antigenic regions of the protein.

It was surprising that human and avian H5 sequences showed substantial differences in their evolutionary patterns, and in particular, that sites in the SABR showed elevated $\omega$ in human H5 compared with avian H5. In contrast to H1 and H3, human and avian H5 sequences cannot be considered distinct viral lineages. Because H5 cannot be transmitted directly among humans, one would expect that human H5 sequences should look like a random sample of avian H5 sequences. Yet this was not the case. In particular, sites in the SABR evolved significantly differently in human and avian H5. There are two possible explanations for this observation: first, the human physiology starts exerting a selection pressure on the SABR as soon as a human is infected with H5, and the H5 sequences evolve significantly within patients. Second, and more likely, only a subset of H5 sequences, in particular sequences with specific changes in the SABR, can infect humans. Therefore, human H5 sequences are not a random sample from avian H5 sequences. Prior work has shown the accumulation of key adaptive substitutions after introduction of avian H1, H2 and H3 to mammalian hosts [31]. The methods we used here are not suitable to identify key substitutions in H5 adaptation from avian to human hosts; however, suitable
A number of prior studies have looked at adaptive evolution in influenza HA using both dN/dS-based methods [33–35] and other methods [14,32,36]. However, a site-by-site comparison across different host species, as we have performed here, is not usually done. One exception is the recent work by Tamuri et al. [36], who identified sites under differential selection pressure in human and avian influenza. Their results are broadly consistent with ours. The majority of sites showed no significantly different selective constraints among human and avian sequences. Nevertheless, a number of sites did show significant differences. For H3, Tamuri et al. [36] identified 11 such sites. This number seems low relative to the large amount of Δω differences between human and avian H3 we found here. Several issues may contribute to this apparent discrepancy: (i) only 11 sites were identified with high confidence by Tamuri et al., and more sites were identified at lower confidence; (ii) we did not attempt to identify for which sites Δω was significantly different among species here; while we saw a substantial amount of variation in Δω, some of this variation may not be statistically significant; (iii) Tamuri et al.’s and our model are fundamentally different and may identify different sites.

To assess whether sites in the SABR had elevated Δω, we used Fisher’s exact test instead of a seemingly more natural linear regression of Δω against RSA and site type. We decided against the regression analysis because Δω is a derived quantity, estimated from a model that used RSA as input data. Therefore, the regression analysis could potentially yield incorrect results. By contrast, for any chosen bisection of the data, Fisher’s exact test answers the question whether SABR sites are enriched on one side of the bisection. Thus, even if the trendlines in figure 2, derived from a regression, are not entirely correct, we can conclude that SABR sites are significantly enriched above the trendlines for all strains except avian H5. One caveat of this application of Fisher’s exact test is that we are treating Δω values as observations, even though they are noisy estimates. The noise on Δω reduces the power of Fisher’s exact test; however, we do not expect the noise to artificially inflate the rate of false positives.

We fitted both RSA-dependent and RSA-independent models to all alignments, and found that in all except one case the RSA-dependent model produced the better fit. Owing to the way, we chose to weight per-site evolutionary-rate estimates by each site’s posterior probabilities, in all cases, estimates obtained by RSA-independent models were highly correlated with Δω estimates obtained by RSA-dependent models (table 1). Hence, rather than explicitly defining an RSA-dependent model, one could make estimates within an RSA-independent framework; then, structural information could be inferred in post-analysis by taking into account partial occupancy of multiple rate categories for each site in the protein. Our analysis shows that incorporating structure in this way produces less accurate rate estimates than when the structure is already known. However, for future studies, this finding also suggests that even when analysis in a structural context is desired, it may be sufficient to first calculate evolutionary rates without structural input, and then correlate these rates with structural features. While rates calculated with structural input are expected to be more accurate, and can reveal aspects of the data that might have remained obscure otherwise [15], there are likely many applications where the saved computational expense of not fitting structure-aware models is worth the small loss in the accuracy of estimates.

Throughout our analysis, we assumed that the protein structure (and hence RSA) remains constant over time and even across HA subtypes. This approximation is reasonable, because HA is a structurally homogeneous protein; available crystal structures show little structural variation. The full-length form of HA, referred to as HA0, consists of over 500 amino acids. Prior to capsid incorporation and viral burst, HA0 is cleaved into two proteins, HA1 and HA2. Comparing all structures available in the protein database (over 100, both pre- and post-cleavage), we found that greater than 90 per cent of the two post-cleavage subunits varied by less than 2 Å RMSD (not shown). Moreover, most of that variation was the result of flexible regions near chain terminations. Several post-cleavage structures exist; they tend to have very different structures near the inter-chain break site of HA0, owing to the additional chain terminations. To eliminate the unpredictable variation near the site of inter-chain termination, we chose to base our RSA calculations on an HA0 crystal structure. Because the cleavage site is in the transmembrane domain of HA, approximately 80 Å away from the SABR, we do not expect any structural variation in this region to impact our results in any substantive way. Finally, we used a high-quality structure from H1 influenza throughout this work, because there were no suitable, full-length H3 or H5 HA0 structures.

The high-evolutionary conservation of HA structures across subtypes and host-species implies that the immune systems of humans and birds do not generally rely on the three-dimensional structure of HA for binding, but only on the specific sequence of amino acids that is overlaid on that structure. If structure itself were important for immunogenicity, we would expect to see large structural deviations as the host immune system pressured influenza to escape neutralization.

This work was supported by NIH grant (no. R01 GM088344) to C.O.W. We thank Oliver Pybus and Tommy Lam for insightful comments.

References


