Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins

Sergey Kryazhimskiy, Jonathan Dushoff,
Georgii A. Bazykin, Joshua B. Plotkin*

Supplementary Information

This PDF file includes

*To whom correspondence should be addressed. E-mail: jplotkin@sas.upenn.edu

1 Computing $\langle t_{\rm ns} \rangle$

In this section we compute the expected time $\langle t_{\rm ns} \rangle$ that elapses between two random consecutive non-synonymous substitutions.

Denote the branches of the phylogenetic tree by small Latin letters, $a, b, c, \ldots$. The set of all branches form a partially ordered set, so that for every pair of branches a and b that are located on the same lineage we can say either that a precedes b (denoted by $a \prec b$), or that a follows b (denoted by $a \succ b$), or that a equals b ($a = b$). Notations $a \preceq b$ and $a \succeq b$ have an obvious meaning. Suppose that substitutions A and B occurred on branches a and b, respectively. Then note that substitution B is consecutive to substitution A if and only if $a \preceq b$. In this case the time between substitutions A and B is then given by expression (1) in the main text.

Let l_a and n_a be the numbers of sites that experienced a synonymous and a non-synonymous substitution on branch a, respectively. If we draw a random non-synonymous substitution, it will be located at branch a with probability proportional to n_a. If we independently draw (with replacement) a pair of non-synonymous substitutions, the probability that they are located at branches a and b, respectively, is proportional to the product $n_a n_b$. If this pair of substitutions happens to be consecutive (i.e., if $a \preceq b$), then the time between these substitutions equals to $t_{\rm ns} = t(a, b)$ given by expression (1) in the main text with probability proportional to $n_a n_b$. Hence, the expected value of $t_{\rm ns}$ is equal to

$$\langle t_{\rm ns} \rangle = \frac{\sum_{a,b:a \preceq b} n_a \; n_b \; t(a,b)}{\sum_{a,b:a \preceq b} n_a \; n_b},$$

where the sums are taken over all pairs of branches that are located on the same lineage and for which, consequently, the order relation is defined.

2 Computing time between consecutive substitutions

We estimated the mean $m(i,j)$ and the standard deviation $\sigma(i,j)$ of the distribution of times between consecutive substitutions for the site pair (i,j) using the following expressions.

$$\begin{aligned} S_{ij} &= \frac{1}{2^{m_{ij}}} \sum_{k=1}^{2^{m_{ij}}} \left| S_{ij}^{(k)} \right| \\ m(i,j) &= \frac{1}{S_{ij}} \frac{1}{2^{m_{ij}}} \sum_{k=1}^{2^{m_{ij}}} \sum_{\pi \in S_{ij}^{(k)}} t_\pi \\ \sigma(i,j) &= \frac{1}{S_{ij}} \frac{1}{2^{m_{ij}}} \sum_{k=1}^{2^{m_{ij}}} \sum_{\pi \in S_{ij}^{(k)}} t_\pi^2 \end{aligned}$$

Here, $\left|S_{ij}^{(k)}\right|$ denotes the number of elements in the set $S_{ij}^{(k)}$, i.e., the number of consecutive substitution pairs at site pair (i,j) found on the phylogenetic tree with the order $O_{ij}^{(k)}$; S_{ij} is the number of consecutive substitutions for the site pair (i,j) averaged over all orders of substitutions. Other notations are as in the main text.

In order to estimate the time measured in years that elapses between the initial and the subsequent substitution in a putatively epistatic pair, we need to calculate the average number of synonymous substitutions that occur in one year in each of the studied proteins. We obtain this proportionality factor by regressing the number of synonymous substitutions that occurred on the phylogeny between the root and a terminal node against the year of isolation of the sequence at the node. The proportionality factors we obtained are 3.99 for H3, 4.41 for H1, 3.28 for N2 and 3.35 for N1.

3 The "fill method"

The null model for our analysis posits that the non-synonymous substitutions occur randomly on the tree. To generate a random distribution of substitutions on the tree we use the "fill method" which distributes the non-synonymous substitution at different sites among branches, while keeping constant the total number of substitutions on each branch $(l_1, l_2, \ldots, l_M)$ as well as the total number of substitutions at each site $\mathbf{s} = (s_1, s_2, \ldots, s_K)$. The total number of substitutions on the tree is then $\sum_k l_k = \sum_i s_i \equiv N$. Suppose that the branches are sorted by length in the descending order, i.e., $l_1 \geq l_2 \geq \ldots \geq l_N$. The fill algorithm works as follows.

1. Initialize the algorithm: set the vector of remaining substitutions $\mathbf{s}' = \mathbf{s}$; let the current branch be $j = 1$. let the set of sites which experienced a substitution at branch j be $\sigma_j = \emptyset$ for all $j = 1, \ldots, M$.

2. Draw a site with the proportional to its weight in the vector $\mathbf{s}'$. Suppose site k was drawn. Check if site k is already in the set j. If it is, then discard the entire run and start from step 1. If it is not, then assign one substitution at this site to branch j, $\sigma_j \to \sigma_j \cup \{k\}$.

3. Update the vector of remaining substitutions $\mathbf{s}' \to \mathbf{s}' - (0, 0, \ldots, \underbrace{-1}_{k\text{-th position}}, \ldots, 0)$.

4. Repeat steps 2 and 3 until branch j is filled, i.e., $|\sigma_j| = l_j$. Update $j \to j + 1$.

5. Repeat steps 2 to 4 until all branches are filled.

This algorithm is guaranteed to sample the space of possible permutations of substitutions among branches uniformly. It does, however, produce a substantial fraction of failed runs (when a site is drawn that is already present in the set σ_j).

4 Uncertainty in phylogeny and ancestral state reconstruction

Here we quantify how uncertainty in the phylogenetic tree and ancestral state reconstruction influences our results. We created 10 bootstrap alignments by randomly re-sampling (with replacement) the positions in our H3 alignment and reconstructed the maximum likelihood phylogenies using these bootstrapped data sets. Given the 10 perturbed phylogenies, we inferred the ancestral states by maximum likelihood and proceeded with our analysis as described in the main text, but generating 1000 permutations per phylogeny and using 100 of these permutations to obtain the FDR and the P-value for the number of positives. The results are presented in Table S4. The results obtained with most of the bootstrap trees were similar to the results reported in the main text, although the significance levels were generally lower than in the original analysis, at least at the nominal level of significance of 0.01. This is presumably caused by inaccuracies of phylogeny reconstruction associated with bootstrapping the alignment. Nevertheless, we found large overlaps between the lists of site pairs implicated in epistasis in the bootstrap analyses and the list of pairs found in our original analysis (see Table S4). Moreover, when we lowered the nominal significance level to 0.005, the number of positives became significant in all bootstrap data sets.

5 Comparison with the BGM method

Here we compare our list of putatively epistatic sites with those by a recently proposed method for detecting co-evolution between sites [S1], implemented in the package HyPhy [S2]. The latter method infers the networks of sites that appear to co-evolve using a Bayesian graphical model (BGM). One of the key differences between the method by Poon et al and the approach suggested here is the way in which timing between substitutions is taken into account. In the method by Poon et al, only those sites that experience substitutions at the same branches of the phylogeny more often than expected, are implicated as epistatic. This implies that the selective differences between the single and the double mutants are so strong that the second substitution follows the first substitution very quickly (see also the main text). By contrast, in our method, sites at which substitutions often follow each other in close temporal succession (but not necessarily fall on the same branches of the phylogeny) are considered to be epistatically interacting.

We applied the BGM analysis implemented in HyPhy to our data sets under the following parameters: no branch corrections while estimate dN/dS values; resolve ambiguities in the ancestral state reconstruction; include only sites with at least 2 substitutions; maximum number of parents: 1; number of MCMC chain iterations: 10^5; number of burn-in steps: 10^4; sample from every 100 steps; no ancestral resampling. We used two types of sequence and phylogeny data sets as input to HyPhy. (a) We input the whole data set of

existing sequences for each protein and the full maximum likelihood (ML) phylogenetic trees that we previously obtained. (b) We removed all terminal nodes from the full ML phylogenetic tree, so that the internal nodes closest to the tips in the original tree became terminal nodes in the truncated tree. We used this truncated tree as well as the inferred sequences at terminal nodes of this tree as input to HyPhy. The complete lists of site pairs with their posterior probabilities of co-evolution are presented in Supplementary Tables S2 and S3.

In order to compare the results of the BGM analysis to our results, we ranked, for each protein, all ordered pairs of sites by their posterior probability of co-evolution, and selected the first n site pairs so that the average posterior probability of co-evolution among them was approximately equal to the estimated FDR rate in our analysis, at the nominal P-value of 0.01. We found $n = 33$ (H3), $n = 40$ (N2), $n = 87$ (H1), and $n = 87$ (N1) pairs in analysis (a) and $n = 33$ (H3), $n = 23$ (N2), $n = 60$ (H1), and $n = 49$ (N1) pairs in analysis (b). We also found that the posterior probability of coevolution between sites 222 and 234 and the drug-resistance site 275 in N1 was smaller than 0.01. These results suggest that our method may have more power in detecting epistasis between sites than the method by Poon et al, at least for the influenza virus surface proteins which evolve rapidly and are well sampled.

References

1. Poon AFY, Lewis FI, Kosakovsky Pond SL, Frost SDW (2007) An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. PLoS Comput Biol 3:e231.

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