Using sequence data to inform antigenic map

June 6, 2019

Data

- Titer data from Bedford et al. (2014)
- HA amino acid sequence data from GISAID and IRD (wasn't able to get one of the sequences so I removed it from titer data as well)

Antigenic map based on sequences

- Align sequences using the DECIPHER package (Wright, 2016). This package performs what they call "profile-to-profile" alignment by constructing a tree and alining with respect to the tree; this process is repeated until convergence and some? subset of the alignments are re-aligned based on sum-of-pairs scores. Repeated until convergence.
- Use Hamming distance to compare sequence in 5 regions: Sa, Sb, Ca1, Ca2, and Cb (Anderson et al., 2018). For each region, the Hamming distance is divided by the length of regions and the proportions are averaged across 5 regions. These distances are multiplied by 20 to reflect 20-dimensional immunological shape space described by Smith et al.; this seems kind of arbitrary to me but let's see what happens.
- Classical MDS based on these distances

First, here are 5 maps based on 5 different regions (Fig.). We see slightly different patterns across regions. Overall, there seems to be a circular pattern in each of them? We see similar (but slightly different geometrically) circular when we use the averaged distance (Fig.). Anderson et al. (2018) used a lot more sequence data to create the same map; we can see similar circular patterns in their analysis.

We want to be able to use this map to inform our antigenic maps based on HI titers. We're going to call these locations μ_i, where each μ_i is a 2-dimensional vector (can be generalized into D dimensions). Then, the location of virus X_i and the location of serum Y_i can be assigned multivariate normal priors:

$$X_i \sim \mathcal{N}(\mu_i, \Sigma_r), Y_i \sim \mathcal{N}(\mu_i, \Sigma_r) \tag{1}$$

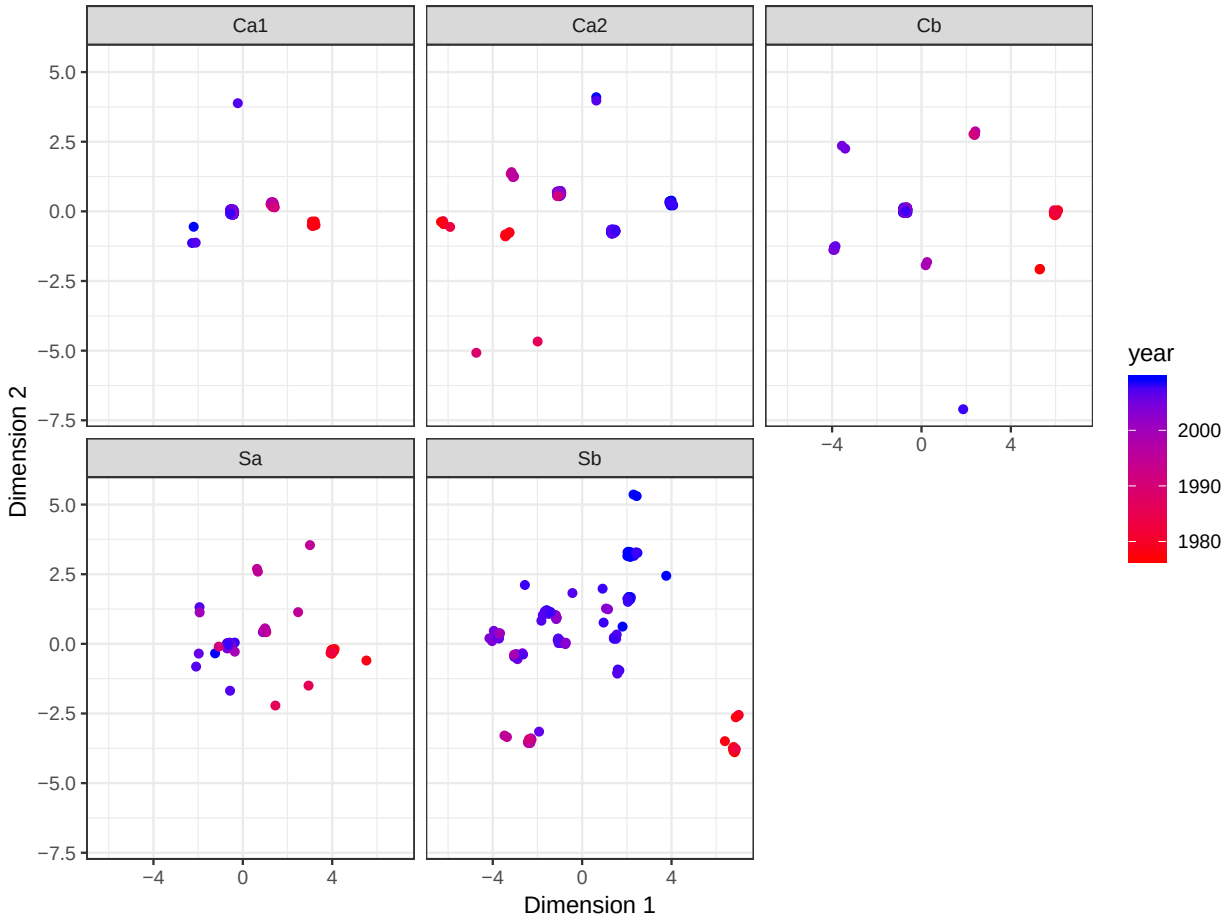

Figure 1: Antigenic map based on Hamming distance (measured as proportion of mutations) across 5 regions. Points are jittered to show overlapping points.

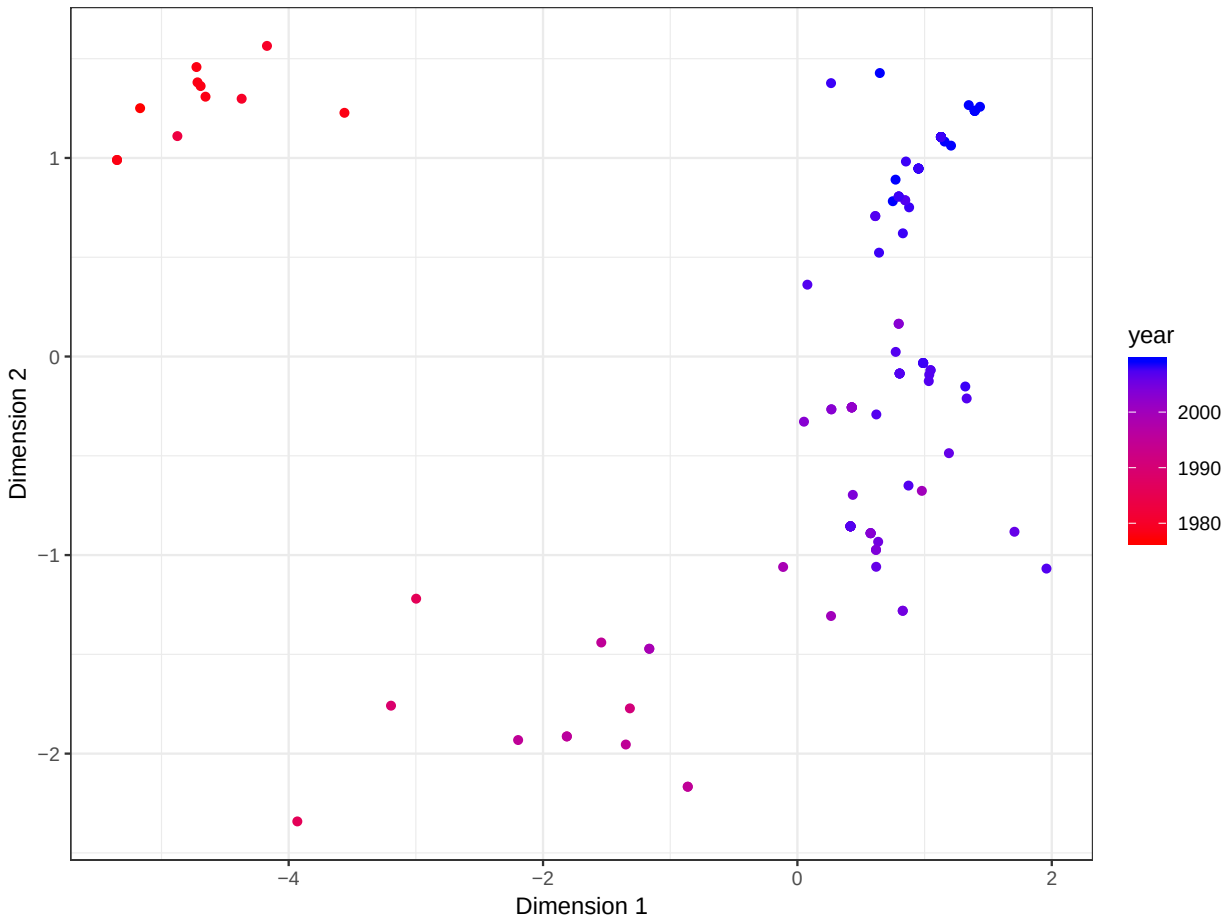

Figure 2: Antigenic map based on mean Hamming distance (measured as proportion of mutations) across 5 regions

where Σ_r is a diagonal matrix whose entries are σ_r^2. I think it's better to think of σ_r^2 as a penalization term rather than trying to find a posterior for it. We might be able to try multiple values of σ_r^2 and do something analogous to cross validation? Not sure yet. For now, I assumed that $\sigma_r^2 = 0.5$ such that the Euclidean distance between a homologous virus and serum will be less than 2 (1.96, technically) *a priori* with 95% probability. This means that homologous pair can have 95% prior probability of having up to 4 fold difference in HI titers with respect to the maximum titer, in the absence of effect of virus or effect of serum. This seems like a biologically reasonable range and a decent amount of penalization? We could try other values later.

Finally, we can model log2 HI titer as usual:

$$HH_{i,j} \sim \mathcal{N}\left(\beta_0 - d(X_i, Y_j) + J_i + A_j, \sigma^2\right), \tag{2}$$

where

$$\begin{aligned} J_i &\sim \mathcal{N}(0, \sigma_J^2) \\ A_j &\sim \mathcal{N}(0, \sigma_A^2) \\ \sigma_J &\sim \text{Gamma}(5, 5) \\ \sigma_A &\sim \text{Gamma}(5, 5) \\ \sigma &\sim \text{Half} - \text{Cauchy}(0, 10) \\ \beta_0 &\sim \mathcal{N}(0, 5) \end{aligned} \tag{3}$$

We should weaken some of these priors later. For now, I'm trying out priors that are not too weak as a preliminary analysis.

Results are summarized in Fig. . We can see that there are noticeable shifts in the location of virus strains. A lot of them don't seem to move around too much but the ones in 1980's seem to move around a lot. Taking HI titres into account seems to preserve the circular patterns (with respect to time) in the map.

See Fig. for effects of virus and antisera. These patterns are very similar to the patterns that we saw before when we didn't use the sequence data. Virus effect seems to decrease over time. There's a weird jump in 2005. Antiserum effect stays constant on average with weird jumps around 2003-2007.

TODO?

- Get more sequences even if we don't have their titre data?
- Better distance metrics based on homology models? Or is hamming distance "good enough"?
- Try to interpret virus effects and serum effects?

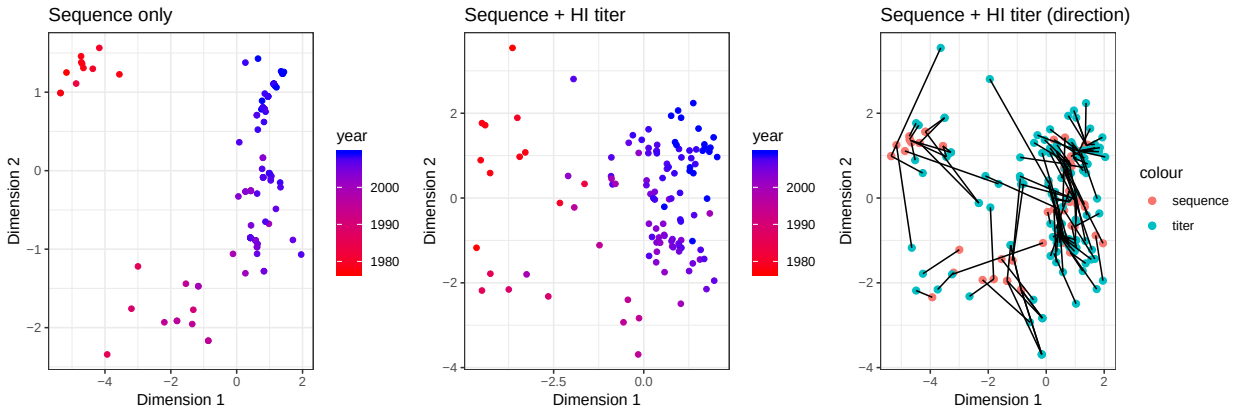

Figure 3: Sequence only: antigenic map based on mean Hamming distance (measured as proportion of mutations) across 5 regions. Sequence + HI titer: antigenic map based on HI titer informed by the sequence map; only showing posterior median. Sequence + HI titer (direction): solid lines connect the virus locations based on the sequence (prior mean) with the locations of the same virus based on HI titres (posterior median).

References

Anderson, C. S., P. R. McCall, H. A. Stern, H. Yang, and D. J. Topham (2018). Antigenic cartography of H1N1 influenza viruses using sequence-based antigenic distance calculation. *BMC bioinformatics 19*(1), 51.

Bedford, T., M. A. Suchard, P. Lemey, G. Dudas, V. Gregory, A. J. Hay, J. W. McCauley, C. A. Russell, D. J. Smith, and A. Rambaut (2014). Integrating influenza antigenic dynamics with molecular evolution. *Elife 3*, e01914.

Wright, E. S. (2016). Using DECIPHER v2. 0 to analyze big biological sequence data in R. *R Journal 8*(1).

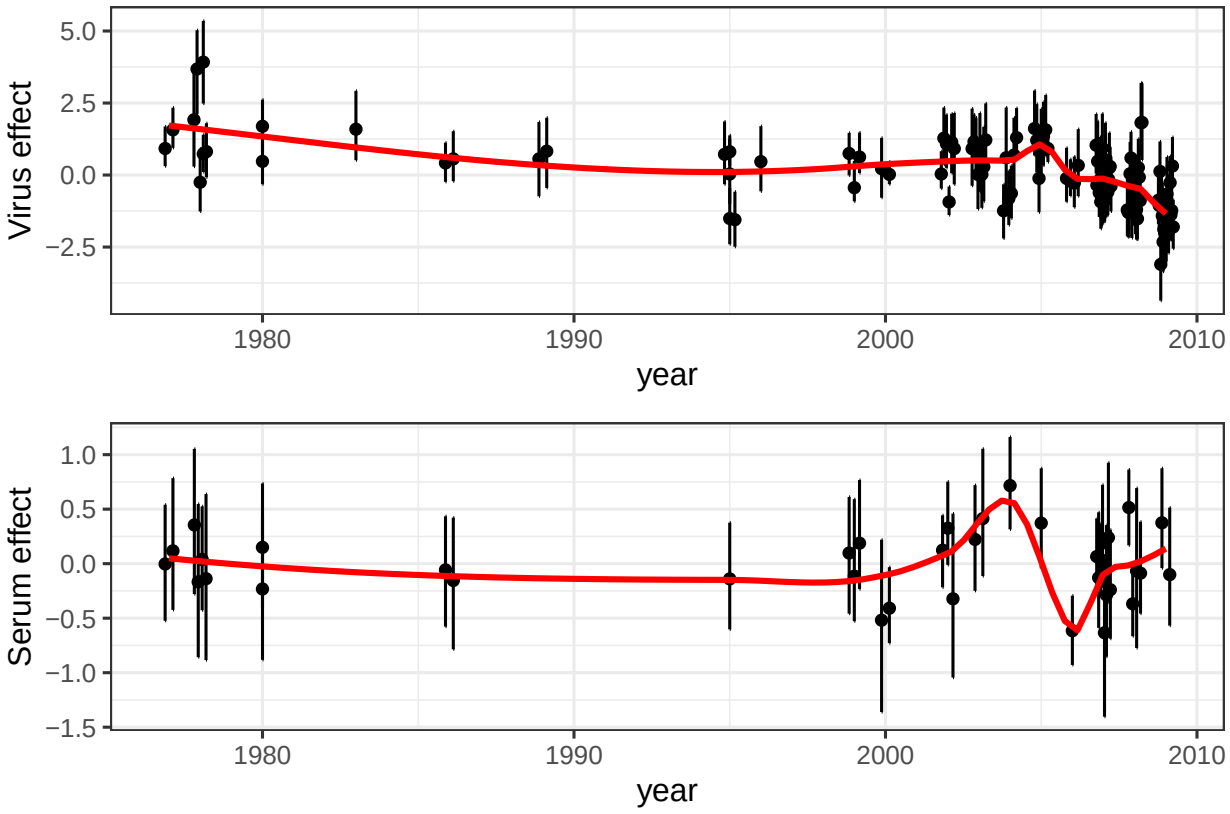

Figure 4: Estimated effects of virus and effects of antisera over time using HI titre data + the sequence data. Showing posterior median and 95% credible intervals.