Project Outline -- a starting point!

In this project we aim to predict which variants of influenza are likely to be successful in the next couple of years. We have established that we can do this using timed phylogenetic trees for influenza - see the preprint at https://www.biorxiv.org/content/10.1101/609248v1

That paper was a first attempt and a kind of proof of principle. There is a lot more to be done! As in any research project, we learned a lot about what we might do next time, and we had new ideas that we didn't have time to test.

So here are some of our thoughts and ideas for the hackseq19 version of flu predictions. If we just got themes 1 and 2 done I'd be quite pleased. If we also, or instead of theme 2, explored some novel approaches to the problem, I think that would be great too.

Theme 1:

Relate the predictions we make to the WHO clade designations and thereby to recent and current flu vaccines. This will require obtaining the WHO clade designation for each sequence in our dataset(s). This, in turn, is done with the help of the nextflu.org's 'augur' pipeline, and a custom phython script. We have included details on the github repo (CladeDesigationNotes)

The results will be essential to answer key questions:

- What would we choose as next year's vaccine?
- What would we have chosen as this year's vaccine?
- Based on what is circulating now, how well would our prediction based on data up to May 2018 have performed as a vaccine?
- What were the vaccines in recent years? How well did they predict which strains were circulating in the following season? For example, did the 2016, 2017, 2018 seasonal vaccines contain strains that were circulating a lot in 2017, 2018 and 2019? In contrast, what would our method have selected in 2016, 2017 and 2018, and did those strains circulate highly in the subsequent years?

  A lot of this last point is simply a matter of google searching (eg what the past vaccines were). How much they were circulating can be obtained to some extent in our data, and to some extent also via literature search.

Theme 2:

Straightforward improvements on the method in the preprint:

- Regression instead of classification
- Explore different subtree selection methods (our default is in the confusingly-named getClades2.R, but it gets *subtrees*, not *clades.* The related file getClades_size.R is a function that uses only the subtree size (without the more complex selection in getClades2). However, this would have to be applied to trees that were sliced up in time, to avoid training on "future" data.

- Proceed as before, but this time, don't bother balancing the groups. Instead, just try to find the few subtrees that actually grow *a lot* in the near future. Here, we could also explore using fewer but larger clades. This would simplify linking our predictions to the WHO designations, but would require using learning methods that adapt for unbalanced group size in the classification (or regression).
- Get epitope information for H1N1 and influenza B so we can use epitope features there. Epitopes are regions in the HA protein that are known to be antigenic; we took information from a paper by Shah et al (cited in the preprint) to identify these for H3N2's HA protein.
- Make epitope features quicker to compute -- reduce the number of pairwise comparisons.
- Make predictions on tips up to 2018.5 and test on the 2019.5 data
  Did we get it right? This links to Theme 1 with the WHO clade designations.

Theme 3: Graph neural networks; other learning methods

Maryam suggested that we could try this task using graph neural networks. We would need to develop ways to avoid passing "future" data in to the training.
A vague outline is as follows:

- The graph is the phylogenetic tree (perhaps terminated at some time T before the present)
- Nodes have features: these could be taken from the local structure of the graph itself in the way we have already done (eg using summary statistics of the phylogeny local to the node). They could be taken from the epitope regions, from other features of the sequence (estimated in the ancestral sequence reconstruction stage in the WHO classification task from Theme 1) at the node and nearby nodes..
- Nodes have to have outcomes - which should be connected to the "success" of the node. Ensuring that we develop an approach that would work *only* seeing data up to a fixed time T (ie training on the part of the graph before a node, not after) would be essential. This seems like a challenge.

If GNNs turn out not to be the way to go, there are other learning methods. For example:

- Make images of the subtrees and use them (this would be a small dataset from the point of view of image classification!)
- Develop an alternative approach that is based on individual tips, rather than subtrees.
- Any other ideas!

Theme 4: Simpler methods

This approach might appeal to those who like modelling. It has some advantages - interpretability being high among them. Here's a sketch:

- Create a tree that is truncated at time *T*.
- Divide (say) the tips in years T-5 to T into "our clades" - these might correspond to the WHO-defined clades in later years. They might not quite, but they'd probably be

somewhat close; perhaps for example clade A2/re would have several of our clades rather than just 1.

- Using branching processes (eg birth-death fitting in R), lineages through time plots, diversity through time plots (could be taken from the sequence data directly for example) - develop simple predictions for which of these major clades is growing in the short term.
- Those keen on ML methods could use Gaussian processes for the short-term forecasts.
- Compare the results to which of these clades actually expands when we look at the tree truncated at time *T+1* or *T+2*.