

Phylogenetics for Predicting Virus Evolution

Anticipating next seasons influenza strains

Paul Wiesemeyer

Problem



HOUSTON,
we have a — CHOO!

- Estimated 3 – 11% of the population catch symptomatic influenza each season [Tokars et al., 2018]
- Influenza evolves rapidly and evades immunity and vaccines
- Epidemics and pandemics caused by RNA recombination events
- Science is slow responding to health issues

Problem

- ~ 8 months from submission to publication of a medical paper [AAMC, 2018]
- ~ 6 months from vaccine strain selection to distribution:



Figure 1: [Bedford, 2015]

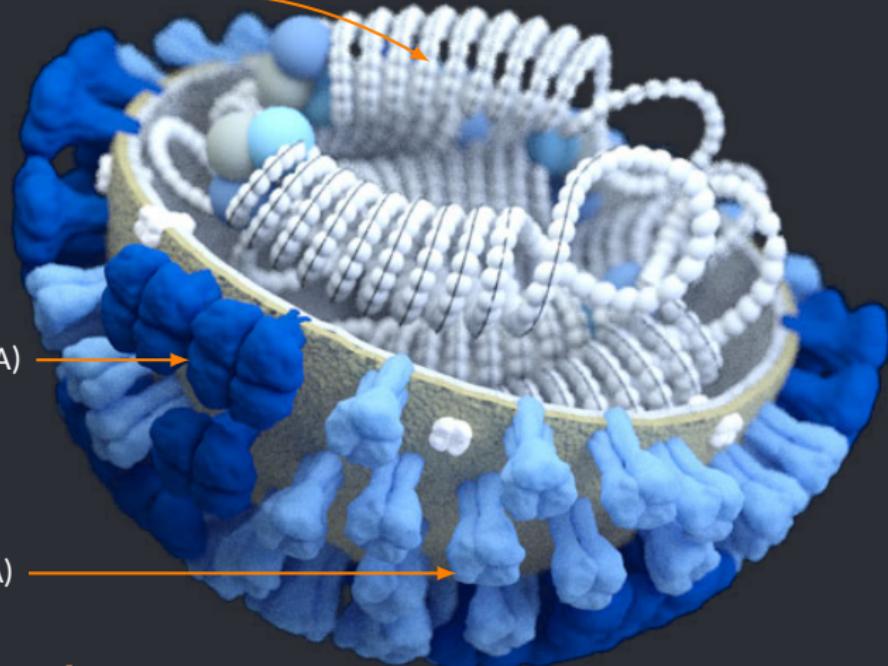
Outline

1. Influenza
2. Phylogenetics
 - 2.1 Sequence Alignment
 - 2.2 The Molecular Clock
3. Predicting the next strain of influenza
 - 3.1 Predicting Virus Evolution
 - 3.2 The Hemagglutination Inhibition Assay
 - 3.3 Mapping Antigenicity to the Tree
 - 3.4 Results
4. *Nextstrain*
 - 4.1 How to use the Framework
 - 4.2 The Powerful Meta Data
 - 4.3 Confidence Levels and Limitations

Influenza

a closer look

8 single stranded RNA



Neuraminidase (NA)

Hemagglutinin (HA)

Image from [CDC, 2020]

Influenza jumps far

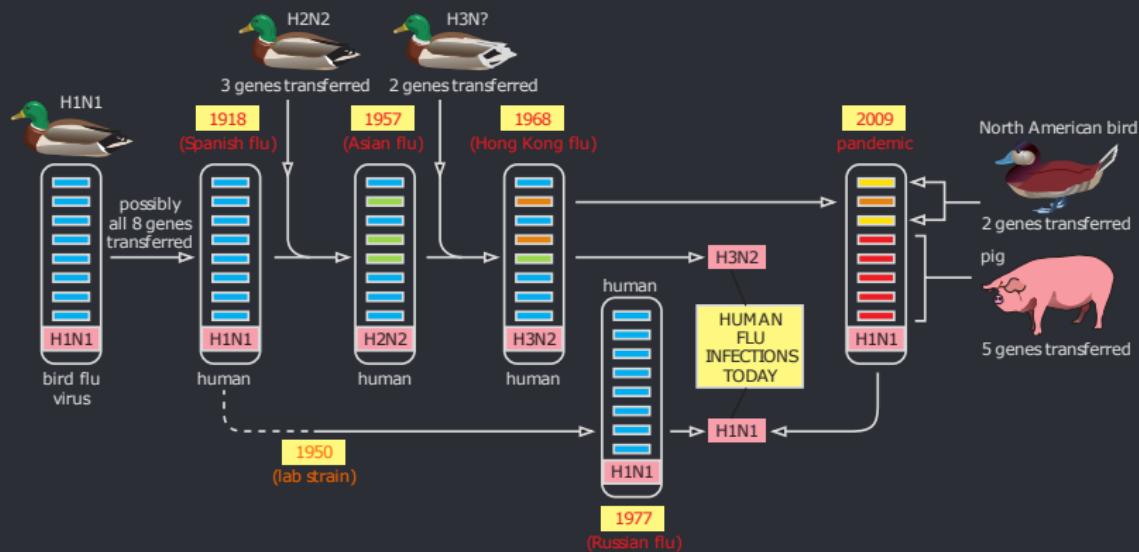


Figure 2: Influenza A evolutionary shift events [Alberts, 2015]

Influenza
runs fast

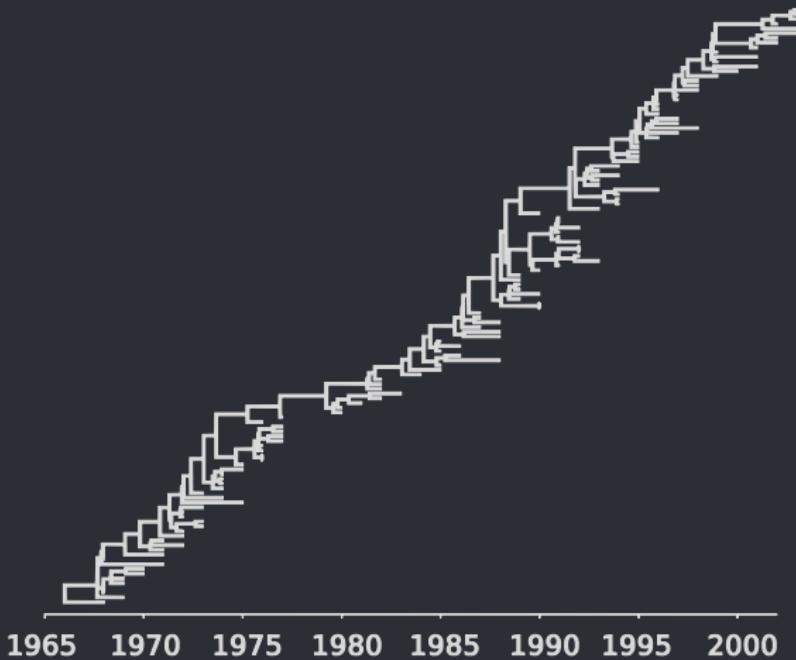


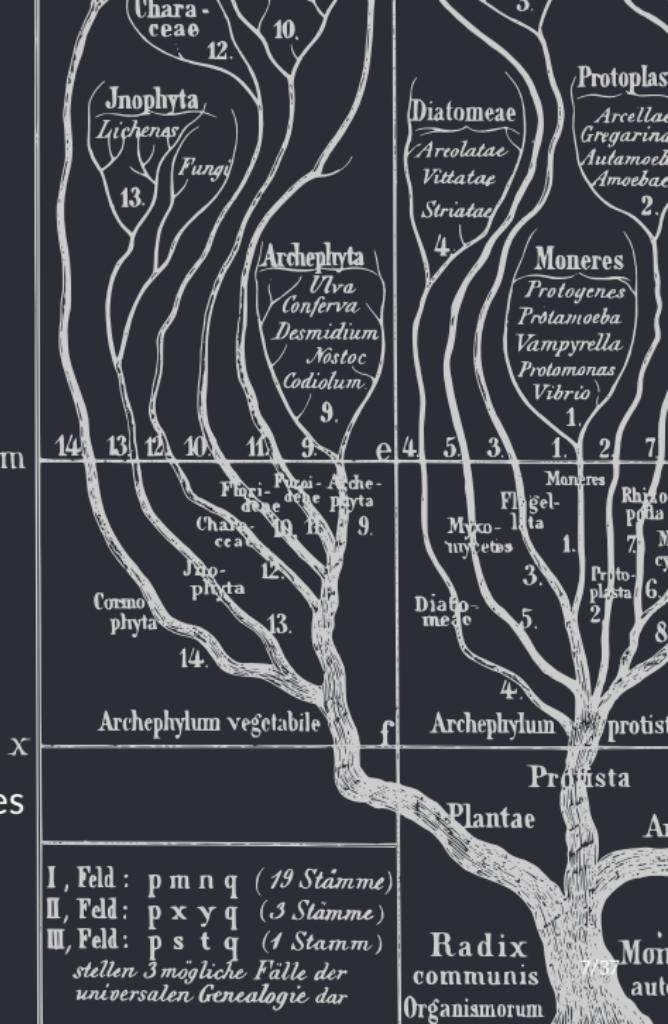
Figure 3: Phylogeny of Influenza A/H3N2 Hemagglutinin [Volz et al., 2013]

Phylogenetics

Terminology

Tree—a simply connected graph

- branches
 - nodes (vertices)
 - leaves (endpoints)
 - root \Rightarrow parents, children
 - clade—a collection of branches



Phylogenetics

Concept: Parsimony

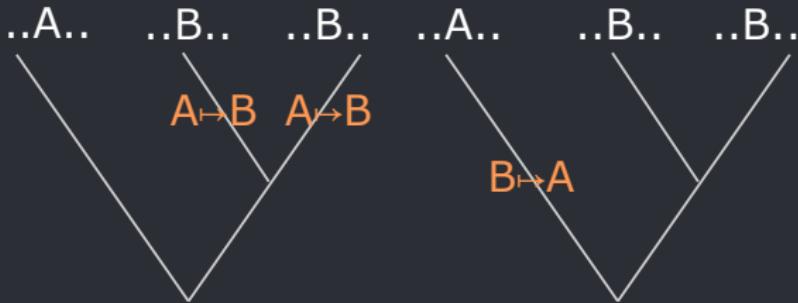


Figure 4: Reconstructing a tree requires likelihood considerations

Minimize the number of mutations to get the most likely tree(s)

Problem: for n leaves, # possible trees $\propto 2^n n!$

⇒ We can always just compute a small number of trees

Phylogenetics

Example algorithm: Sequence Alignment

	A	B	C	D	E	F
A		9	2	4	9	10
B			9	6	2	10
C				5	9	10
D					6	10
E						10
F						

N sequences of lengths $n_1 \dots n_N$ (not same due to indels, seq. errors)

- Pairwise matching, give a penalty, make $N \times N$ table
- Lowest penalties group together, form new subunit.
- Iterate. When comparing to sets of sequences, use mean.

Result: small number of locally most likely trees, not necessarily global maximum.

Effectively in use today: Markov Chain Monte Carlo "climber"

Phylogenetics

Concept: *The Molecular Clock*

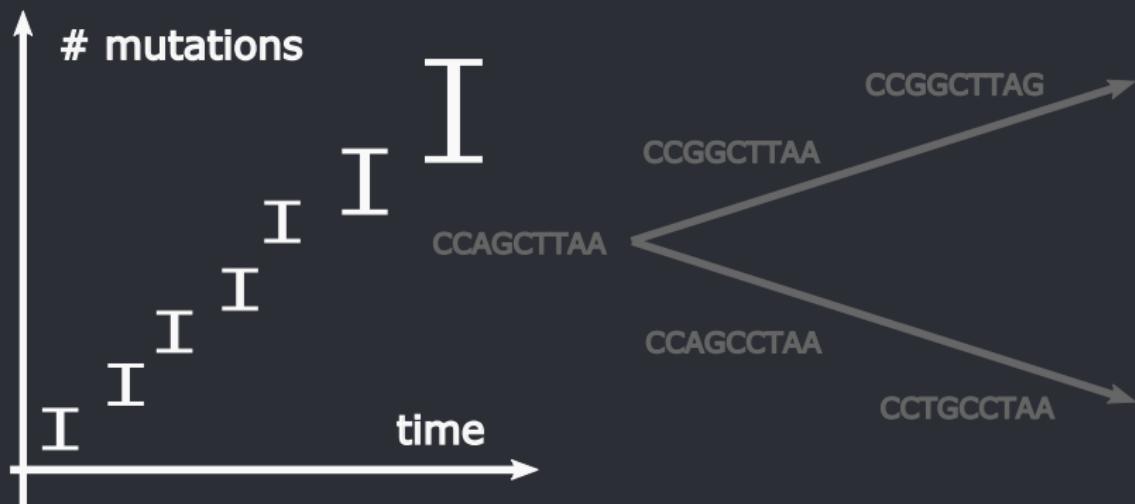


Figure 5: Linear time-mutation relationship

Phylogenetics

Concept: The Molecular Clock

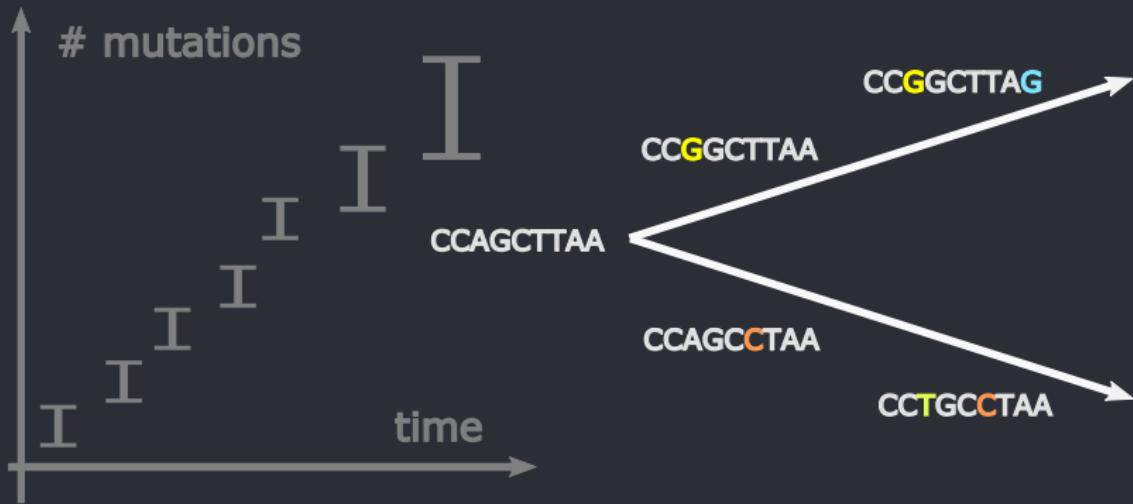
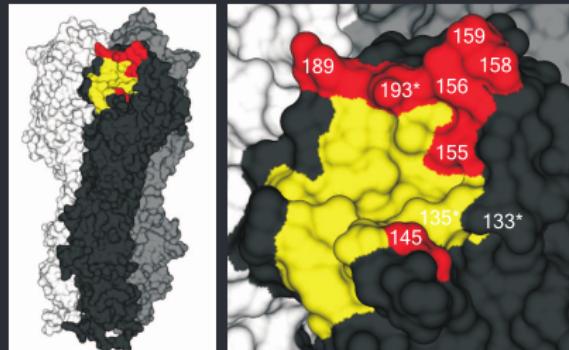
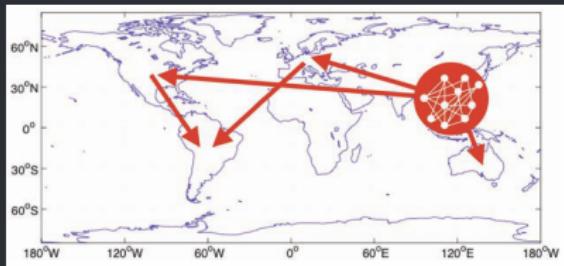


Figure 5: Linear time-mutation relationship

Strongly simplified. Different genome sites experience different evolutionary pressure. Say: Silent vs. non-silent, each site in genome can have a different rate. Then it is hard to deduce anything, possible intermediate: Distribution of rates.

Predicting the next strain of influenza

Approaches



Epidemiology: Geographical region?

Molecular Biology: Is there a certain type of Mutation?

Immunology: Hemagglutinin inhibition cartography

Vision: Bring these levels together

The Hemagglutination Inhibition Assay

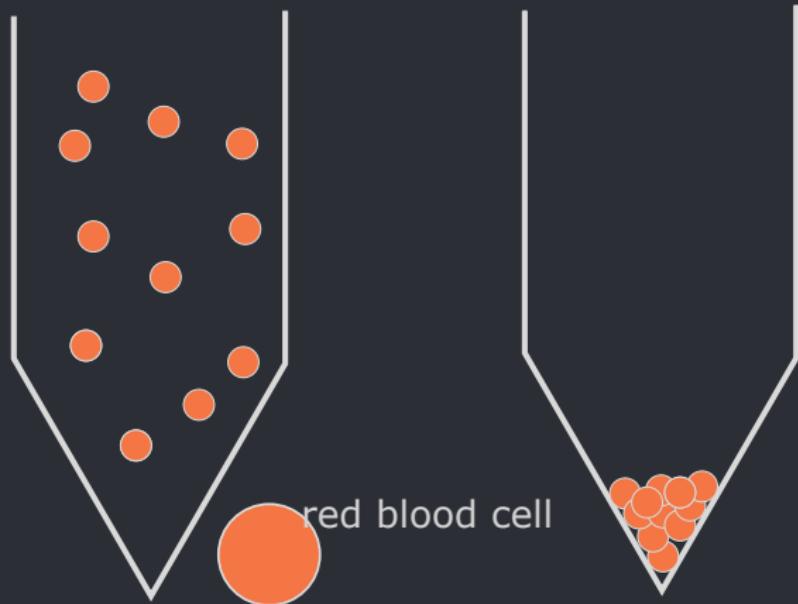


Figure 6: Red blood cells (RBC) precipitate.

The Hemagglutination Inhibition Assay

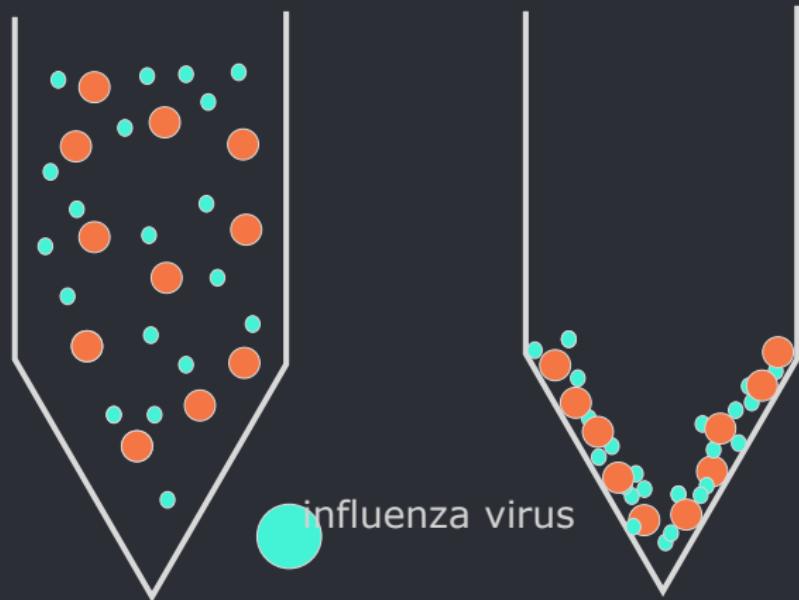


Figure 7: Influenza Hemagglutinin (HA) coagulates the RBC, forming a mat.

The Hemagglutination Inhibition Assay

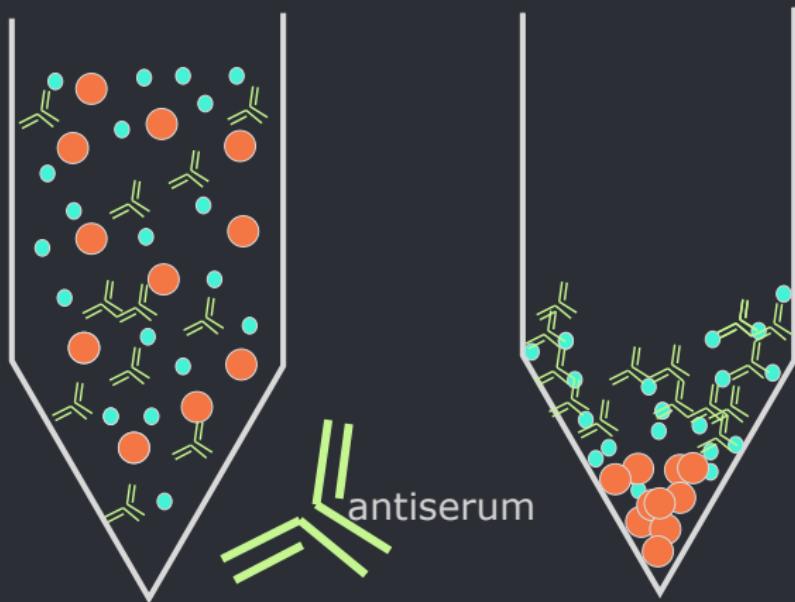


Figure 8: Antisera that fit the HA's epitope site bind it, letting the RBC sink to the bottom. Effect works up to a certain antigenic distance and .

The Hemagglutination Inhibition Assay

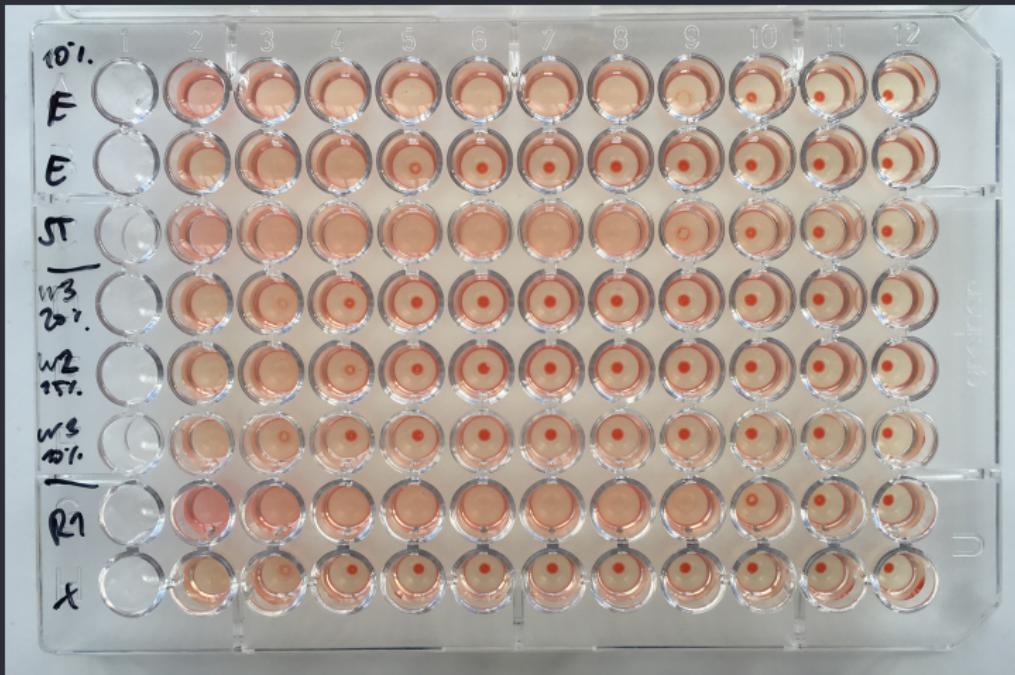
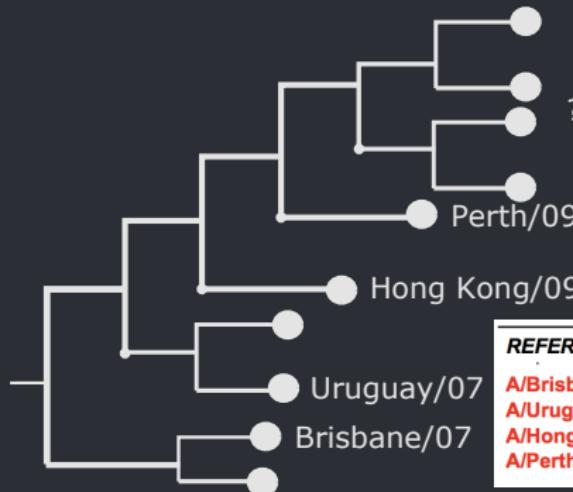


Figure 9: Here, one antiserum is tested in 12 different dilutions against 8 different virus strains. The highest dilution that prevents agglutination is called the titer.

The Hemagglutination Inhibition Assay

How this used to be looked at

Mapping Antigenicity to the Tree



REFERENCE VIRUSES	A/Bris	A/Uru	A/HK	A/Perth
A/Brisbane/10/2007	2560	2560	80	<
A/Uruguay/716/2007	1280	2560	<	<
A/Hong Kong/1985/2009	80	160	1280	640
A/Perth/16/2009	<	40	640	640

Figure 10: Mapping the chart to the tree constructed from sequences

Mapping Antigenicity to the Tree

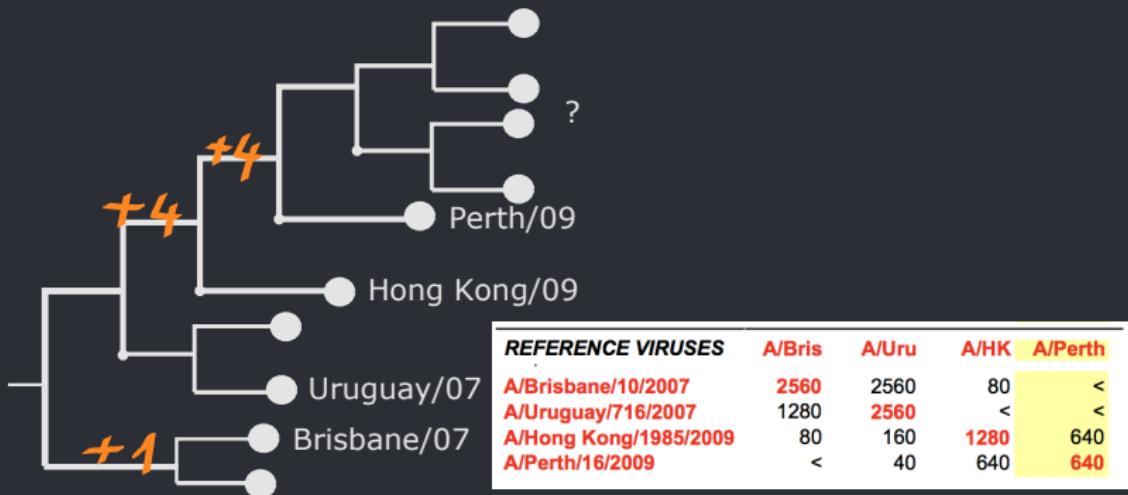
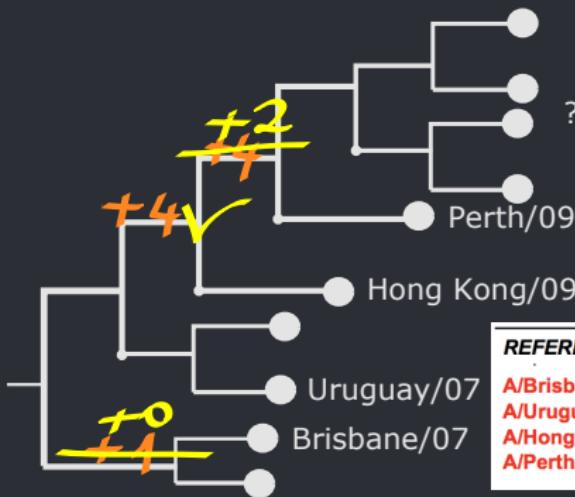


Figure 11: Here, one antiserum is tested in 12 different dilutions against 8 different virus strains. The highest dilution that prevents agglutination is called the titer.

Mapping Antigenicity to the Tree



REFERENCE VIRUSES	A/Bris	A/Uru	A/HK	A/Perth
A/Brisbane/10/2007	2560	2560	80	<
A/Uruguay/716/2007	1280	2560	<	<
A/Hong Kong/1985/2009	80	160	1280	640
A/Perth/16/2009	<	40	640	640

Figure 12: Here, one antiserum is tested in 12 different dilutions against 8 different virus strains. The highest dilution that prevents agglutination is called the titer.

Mapping Antigenicity to the Tree

The asymmetry makes sense, think of it like this but with more dimensions



Figure 13: [rosipaw, 2010]

Formulas I

$T_{a\beta}$ HI titer of virus a against antiserum β (virus b)

$H_{a\beta}$ \log_2 relative titer (we'll use this one)

$$H_{a\beta} = \log_2(T_{b\beta}) - \log_2(T_{a\beta}) \quad (1)$$

$\hat{H}_{a\beta}$ predicted \log_2 relative titer

v_a avidity of virus a (=greediness)

p_β potency of antiserum β (=effectiveness)

$D_{a\beta}$ genetic component of titer drop

$$\hat{H}_{a\beta} = v_a + p_\beta + D_{ab} \quad D \text{ is between two viruses} \quad (2)$$

Formulas II

What remains is split up into a sum over individual mutation contributions:

$$D_{ab} = \sum_{i \in (a \dots b)} d_i \quad (3)$$

Where the sum is over the path connecting virus a and virus $b =$ antiserum β . Now we want

$$\hat{H}_{a\beta} \stackrel{!}{=} H_{a\beta} \quad (4)$$

to that end we minimize a cost function C of the whole tree:

$$C := \sum_{a,\beta} (\hat{H}_{a\beta} - H_{a\beta})^2 + \lambda \sum_i d_i + \gamma \sum_a v_a^2 + \delta \sum_\alpha p_\alpha^2 \quad (5)$$

Results

Omitting 10% of the HI titer data, to make a test

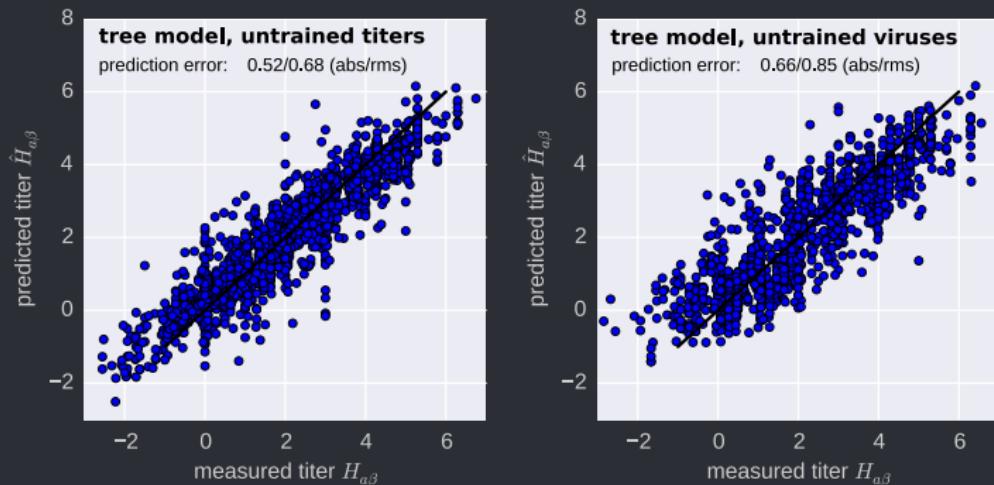


Figure 14: On the left, 10% randomly picked measurements were omitted, on the right, 10 % of entire titer columns were held back, as if it was a new clade

Results

Asymmetry not interesting, but maybe quartet rule?

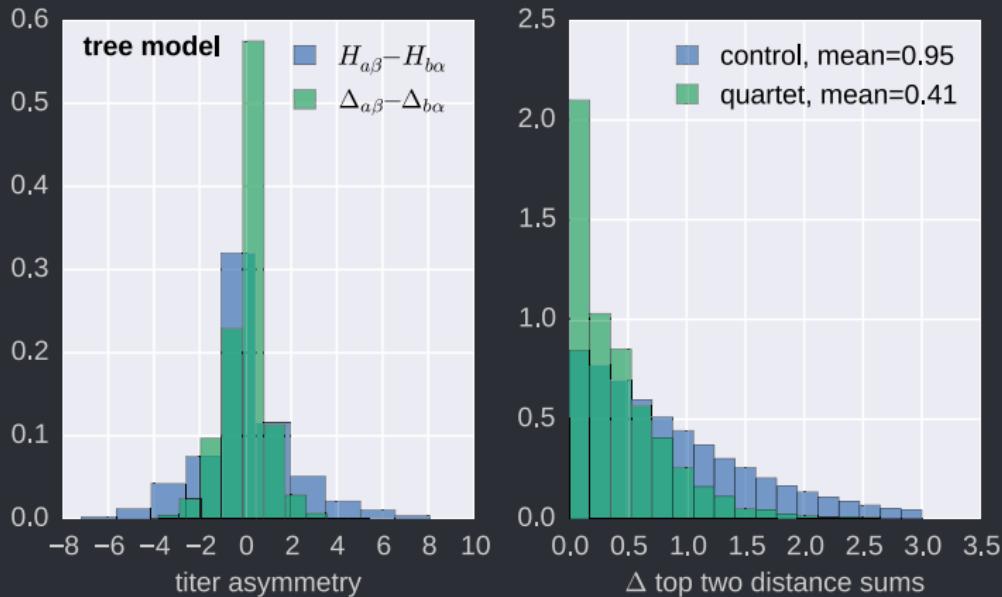


Figure 15: a

Results

Does recent antigenic evolution have fixation implications?

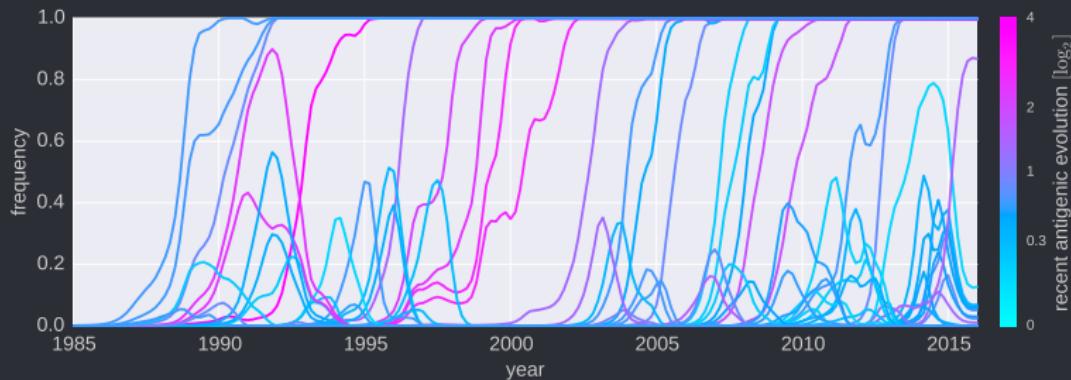


Figure 16: Fraction of samples having a certain mutation plotted vs. time
Strains with frequencies smaller than 0.01 were omitted.

Results

thresholds

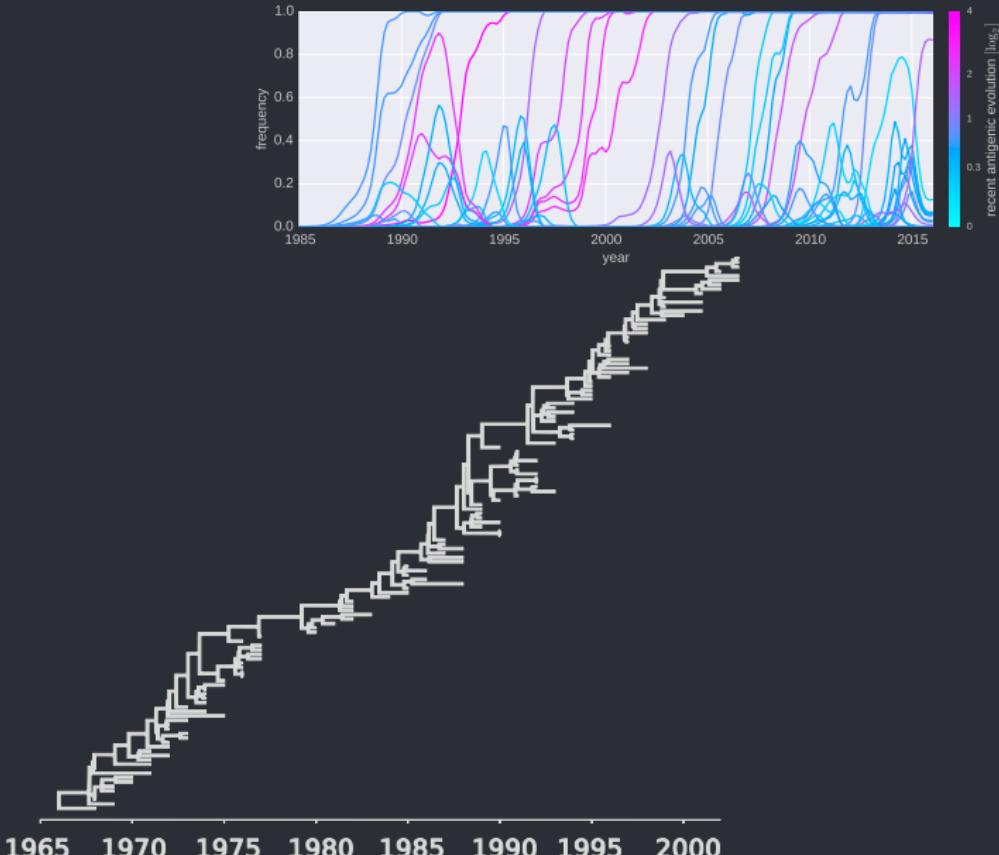


Figure 17: GET nextstrain tree with trunk colored according to antigenic advance

Results

thresholds

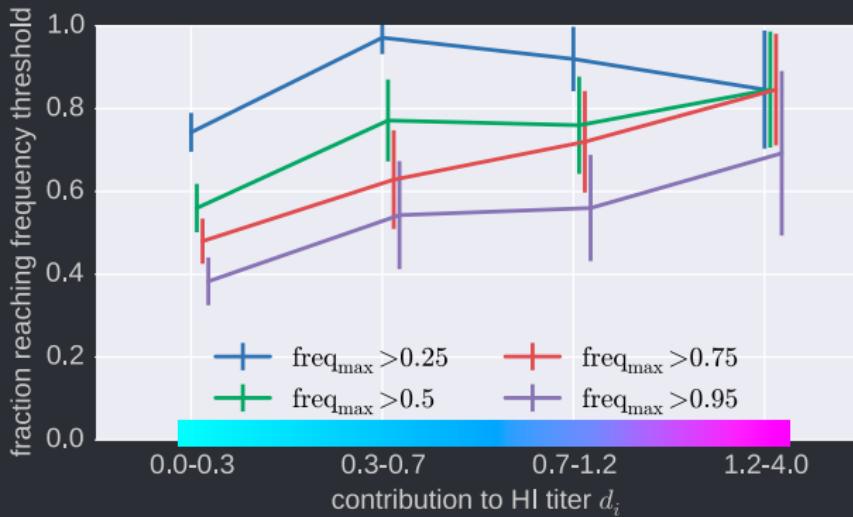
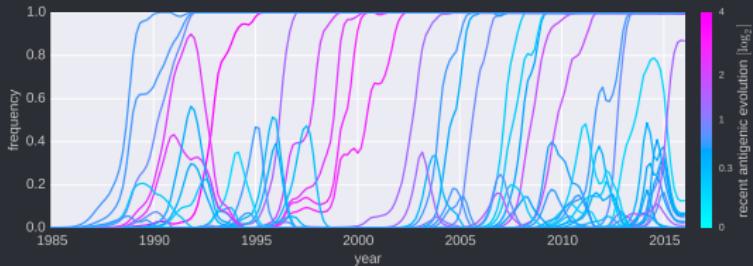


Figure 18: a

Results

distance to season year

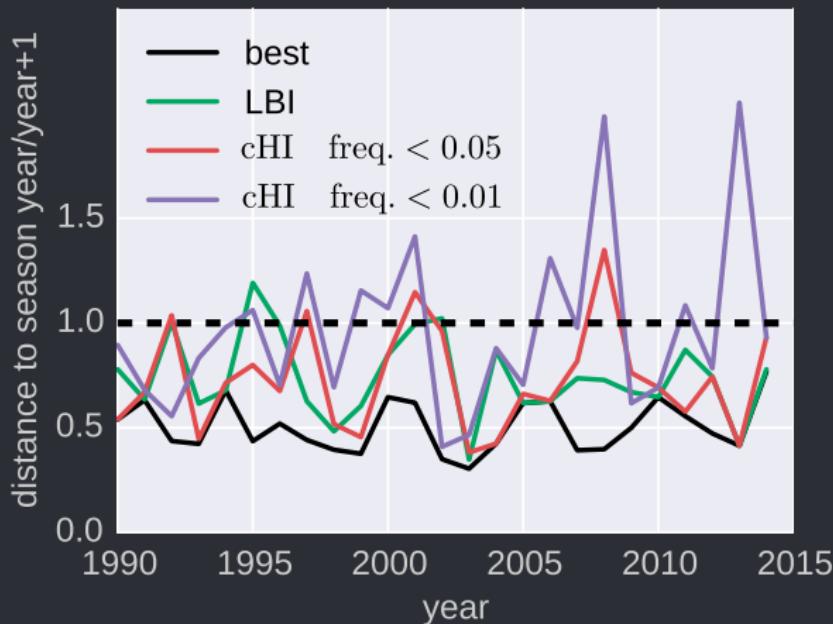


Figure 19: a

Nextstrain

Mending pieces together

Please visit [nextstrain.org/narratives/.....](https://nextstrain.org/narratives/)

Ideas: COVID-Line? H3N2-Line?

Multiple data sets?

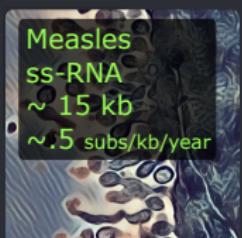
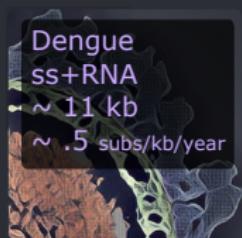
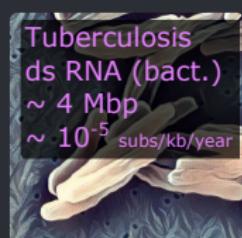
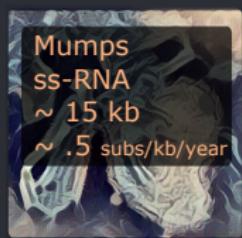
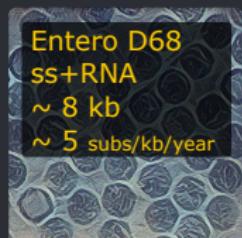
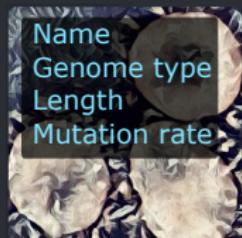
de Vries Epitope sites

Koel 7 sites

D614G — Epidemiologic turning point

Nextstrain: How to use the Framework

<https://nextstrain.org/community/narratives/gitchhiker/virophyle>



Nextstrain: The Powerful Meta Data

(this is mirrored in the narrative)

Nextstrain: Confidence Levels and Limitations

(this is mirrored in the narrative)

Multi Scale Evolution

If a single event mutation occurs, say **D 186 G** in the HA genome, it is subject to multi scale evolutionary selection:

- this RNA instance vs. the other RNA strands in the same cell
- this cell's mutated viruses vs. other viruses inside the host
- this host's viruses vs. viruses in rest of the population
- this population vs. other populations

These scales are difficult to separate. At the population level *epidemiological* processes may dominate.

Outlook

and closing remarks

- Difficult to disentangle levels, therefore integrate visualization
- Include more meta data (symptoms, severity, & c.)
- Interesting: larger evolutionary timescale
- Include other pathogens, like bacteria, CAVEAT: HGT
- My question: what happens if epidemic threshold is crossed even further, and zoonotic forcing is stronger?
- nice to see instance of Red Queen Hypothesis
- Transparency and Community building very important!
- When you find literature on phylodynamics, often animals, but to me this is more difficult.

References I

- [AAMC, 2018] AAMC (2018).
Journal of the Association of American Medical Colleges Frequently Asked Questions.
- [Alberts, 2015] Alberts, B. (2015).
Molecular biology of the cell.
- [Bedford, 2015] Bedford, T. (2015).
Real-time evolutionary forecasting for influenza vaccine strain selection.
- [CDC, 2020] CDC (2020).
Images of Influenza Viruses | CDC.
Library Catalog: www.cdc.gov.
- [Haeckel, 1866] Haeckel, E. (1866).
Generelle Morphologie der Organismen. Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Descendenz-Theorie, etc, volume 2.
- [Koel et al., 2013] Koel, B. F., Burke, D. F., Bestebroer, T. M., Vliet, S. v. d., Zondag, G. C. M., Vervaet, G., Skepner, E., Lewis, N. S., Spronken, M. I. J., Russell, C. A., Eropkin, M. Y., Hurt, A. C., Barr, I. G., Jong, J. C. d., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., Fouchier, R. A. M., and Smith, D. J. (2013).
Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution.
Science, 342(6161):976–979.
Publisher: American Association for the Advancement of Science Section: Report.
- [rosipaw, 2010] rosipaw (2010).
146/365 square peg into a round hole.

References II

- [Russell et al., 2008] Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., Jong, J. C. d., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D. M. E., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A. M., and Smith, D. J. (2008).
The Global Circulation of Seasonal Influenza A (H3N2) Viruses.
Science, 320(5874):340–346.
Publisher: American Association for the Advancement of Science Section: Research Article.
- [Tokars et al., 2018] Tokars, J. I., Olsen, S. J., and Reed, C. (2018).
Seasonal Incidence of Symptomatic Influenza in the United States.
Clinical Infectious Diseases, 66(10):1511–1518.
Publisher: Oxford Academic.
- [Volz et al., 2013] Volz, E. M., Koelle, K., and Bedford, T. (2013).
Viral Phylodynamics.
PLoS Computational Biology, 9(3).